

# Context-Dependent Kernels for Object Classification

Hichem Sahbi, Jean-Yves Audibert and Renaud Keriven

**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 machines (SVMs). In their standard version, commonly used kernels such as the Gaussian one, 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.

We focus in this paper on object recognition using a new type of kernel referred to as “context-dependent”. Objects, seen as constellations of interest points are matched by minimizing an energy function mixing (1) a fidelity term which measures the quality of feature matching, (2) a neighborhood criterion which captures the object geometry and (3) a regularization term. We will show that the fixed-point of this energy is a context-dependent kernel (CDK) which is also positive definite. Experiments conducted on object recognition show that when plugging our kernel in SVMs, we clearly outperform SVMs with context-free kernels (CFK).

**Index Terms**—Kernel Design, Statistical Machine Learning, Support Vector Machines, Context-Free Kernels, Context-Dependent Kernels, Object Recognition.



## 1 INTRODUCTION

INITIALLY introduced in [5], kernel methods including support vector machines (SVMs) show a particular interest as they are performant and theoretically well grounded [26]. These methods rely on the hypothesis of the existence of (explicit or implicit) functions which map training and test data from *input* spaces into high dimensional Hilbert spaces (see for instance [29]). Kernels are symmetric, continuous, bi-variate similarity functions which take high values when input data share similar structures or appearances 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 and regardless their pose. A wide range of vision applications are tackled using kernel methods including optical character recognition [20], pose estimation [22], image retrieval [33] and the most studied object recognition problem [19], [2], [12]. In almost all the proposed solutions, authors use and combine, via algebraic operations, standard kernels such as the linear, the polynomial and the Gaussian

[11]. These kernels also referred to as *holistic* are defined on fixed length and ordered data [8], i.e., into Euclidean spaces including color, shape and texture spaces [32]. Even though proved to be relatively performant, holistic kernels lack a priori knowledge about the application tasks and the expected properties of invariance (such as linear and non-linear transformations.)

A second generation of kernels, referred to as *local*, has recently emerged as an alternative to holistic ones. Local kernels are defined on structured data [10], i.e., which cannot be represented in fixed length and ordered spaces, such as interest points, regions, graphs, trees, etc. Both holistic and local kernels should satisfy certain properties among them the positive definiteness, and preferred kernels should have 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 ones. While the design of kernels gathering flexibility, invariance and low complexity is a challenging task; the proof of their positive definiteness (PD) is sometimes harder [9]. PD may ensure, according to Vapnik’s SVM theory [34], optimal *theoretical* generalization performance and also the uniqueness of the SVM solution [5].

Considering a database of objects (images), each one represented by a vector set, for instance a constellation of interest points [25], [18], extracted using any suitable corner detector [13]. Two families of local kernels can be found in the literature, in order to handle this type of data; those based on statistical “length and order insensitive” measures such as the Kullback Leibler

• H. Sahbi is with the CNRS LTCL, Télécom ParisTech, 46 rue Barrault, 75013 Paris, France.

E-mail: hichem.sahbi@telecom-paristech.fr

• J.-Y. Audibert is with Université Paris-Est, LIGM, Imagine, 6 avenue Blaise Pascal, 77455 Marne-la-Vallée, France, & Willow, CNRS/ENS/INRIA — UMR 8548, 23 avenue d’Italie, 75214 Paris, France. E-mail: audibert@imagine.enpc.fr

• R. Keriven is with Université Paris-Est, LIGM, Imagine, 6 avenue Blaise Pascal, 77455 Marne-la-Vallée, France. E-mail: keriven@imagine.enpc.fr

Manuscript received 07 Sep. 2009; revised 02 Aug. 2010; accepted 05 Oct. 2010.

divergence, and those which require a preliminary step of alignment. In the first family of local kernels, the authors in [16], [21] estimate for each object (set of vectors) a probability distribution and compute the kernel between two objects (two distributions) using the “Kullback Leibler divergence” in [21] and the “Bhattacharyya affinity” in [16]. Only the kernel in [16] satisfies the PD condition and both kernels were applied for image recognition tasks. In [36], the authors discuss a new type of kernel referred to as “principal angles” which is PD. Its definition is based on the computation of the principal angles between two linear subspaces under an orthogonality constraint. The authors demonstrate the effectiveness 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 [28]. The main drawback, of this first family of local kernel, resides in the strong assumptions about the used probabilistic functions in order to model the distributions of the sets of vectors, as these assumptions may not hold true in practice.

In the second family of local kernels, the one in [35] (called the “max”) considers the similarity, between two vector sets, as the sum of their highest matching scores and unlike discussed in [35] this kernel is actually not PD [1]. In [19], the authors introduced the “circular-shift” kernel defined as a weighted combination of PD kernels using an exponent. The latter is chosen in order to give more prominence to the largest terms so the resulting similarity approximates the “max” and also satisfies the PD condition. The authors combined interest points 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 [6], the authors introduced the “intermediate” matching kernel, for object recognition, which uses virtual interest points in order to approximate the “max” while satisfying the PD condition. Recently, [12] introduced the “pyramid-match” kernel, for object recognition and document analysis, which maps interest points using a multi-resolution histogram representation and computes the similarity using a weighted histogram intersection. The authors showed that their kernel is PD and can be computed linearly with respect to the number of interest points. Other matching kernels include the “dynamic time warping” kernel which provides, in [1], an effective matching strategy for handwritten character recognition, nevertheless the PD condition is not guaranteed.

### 1.1 Motivation

The success of the second family of local kernels strongly depends on the quality of interest point alignments mainly when images contain repeatable and redundant structures. Regardless the PD condition, a *naive* matching kernel, i.e., which given two objects, looks for all pairs

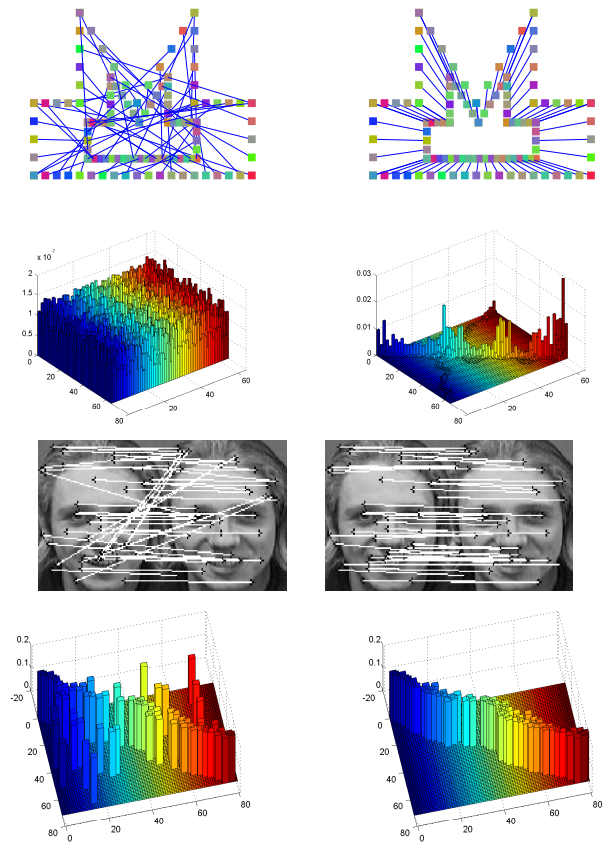


Fig. 1. Considering the two (nested) subsets of interest points shown in row 1 (denoted  $\mathcal{S}_p, \mathcal{S}_q$  and their indices  $\mathcal{I}_p = \{1, \dots, n\}, \mathcal{I}_q = \{1, \dots, m\}$ ).

**The first and the third rows:** show matching pairs defined as  $\{(x_i^p, x_j^q), x_j^q = \arg \max_{x_\ell^q} k(x_i^p, x_\ell^q), x_i^p \in \mathcal{S}_p, x_\ell^q \in \mathcal{S}_q\}$ , where  $k$  denotes either a context-free kernel, actually the Gaussian (left) or our CDK (right). The 3D RGB attributes of interest points in  $\mathcal{S}_q$  correspond to random perturbations (additive uniform noise) of the 3D RGB attributes of  $\mathcal{S}_p$ .

**The second and the fourth rows:** show the values of  $k(x_i^p, x_j^q), i \in \mathcal{I}_p, j \in \mathcal{I}_q$  using a context-free kernel (left) and our CDK (right). Colors are used only for ease of visualization and the x-y axis labels respectively correspond to the indices in  $\mathcal{I}_p$  and  $\mathcal{I}_q$ . We clearly see, through the right-hand sides of rows 2 and 4, that the highest values of  $k$  correspond to the correct matches.

of interest point similarities, using a context<sup>1</sup>-free kernel (such as the Gaussian), and sums the largest similarities, might result into many false matches. Fig. (1, left) illustrates the deficiency of context-free kernels when estimating the matching and also similarity between two groups of interest points. The Gaussian kernel is used in order to evaluate this similarity matrix, between all the pairs of interest points, each one represented by its 3D RGB color attributes. Any slight perturbation of the values 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 [25],

1. Given a set of interest points  $\mathcal{X}$ , the context of  $x \in \mathcal{X}$  is defined as the set of points spatially close to  $x$  and with some particular geometrical constraints (see section 2.2 and also [24] for a detailed and a formal definition of the context.)

[4], [30], for the general problem of visual matching, about the strong spatial and geometric correlation and distortion between interest points in the image space. This limitation also appears in closely related areas, in machine learning, 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 also [24]). 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, a “context” criterion and an “entropy” term. The fidelity term is inversely proportional to the expectation of the Euclidean distance between the most likely aligned interest points while the context criterion measures the spatial coherence of the alignments, i.e., how good two interest points, with exactly the same context, match. Given a pair of interest points  $(f_p, f_q)$  with a high alignment score (defined by kernel value), the context criterion is proportional to the alignment scores of all the pairs close to  $(f_p, f_q)$  but with a given spatial configuration. The “entropy” term considers that without any a priori knowledge about the alignment scores between pairs of interest points, the joint probability distribution related to these scores should be as flat as possible so this term acts as a *regularizer*.

The general form of CDK captures the similarity between any two interest points by incorporating their context, i.e., the similarity of the interest points with exactly the same spatial configuration with respect to  $(f_p, f_q)$ . Our proposed kernel belongs to the same family as the “dynamic time warping (DTW)” kernel [1]. The latter is based on the Viterbi alignment distance, between two sequences of vectors, that allows forward steps of size one in one of the two sequences or both of them; this is commonly known as the “ordering constraint”. We consider instead a “context constraint” (also referred to as “neighborhood constraint”) which states that two points match if they have similar features<sup>2</sup> and if their neighbors, with exactly the same spatial geometric configuration, match too. This also appears in other well studied kernels such as Fisher [15], which implements the conditional dependency between data using the

Markov assumption. CDK implements such dependency while also 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 and discrimination by allowing each interest point to consider its context in the matching process. Moreover, the proposed alignment method (and hence our kernel design) is model-free, i.e., it is not based on any a priori alignment model such as homography which might not capture the actual inter-object transformations; for instance when objects deform. Even though we investigated CDKs in the particular task of object recognition, we can easily extend it to handle closely related areas in machine learning such as text alignment for document retrieval [23], machine translation [31] and bioinformatics [27].

Given a database of images, each one seen as a constellation of interest points. We use convolution kernels [14] in order to build Gram matrices, consisting of all the possible cross similarities of images in the database. A convolution kernel, as will be reminded in Section (2.1), is the sum of all possible cross similarities each one computed using a *minor kernel*; actually CDK. Notice that this work is the continuation of [24] with several updates:

- The definition of the context is updated using finer statistics about the co-occurrences of interest points at different orientations and locations (see Section 2.2).
- The theoretical results about the positive definiteness and the convergence to the fixed point are updated and now our framework provides better and loose constraints about the setting of the context weight parameter defined in Section 2.4.
- And, as will be shown in proposition 2, a Gram matrix built using CDK, on any finite set with no duplicate data, will be full rank so invertible.

In case of SVM the last property guarantees, for any training set whatever its labeling, the existence of a separating classifier. 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 [34] infinite. Nevertheless, this dimension is bounded by the size of the training set as SVMs find the solution in the span of training data so theoretical generalization bounds [34] are in practice not loose.

Notice that the setting of CDK is transductive, i.e., the training and also the *whole* testing data should be available in order to train and apply SVM. For some applications this might limit the usability of CDK mainly for sequential testing (i.e., when test data comes sequentially in time), but it can still be used for many batch testing applications as shown in experiments.

We consider the following organization of the paper; we first introduce in Section 2, our energy function

<sup>2</sup>. In practice, we define features as coefficients of SIFT [18], shape context [3] or self similarity [30].

which makes it possible to design our context-dependent kernel and we show that this kernel is positive definite 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. We conclude in Section 5 and provide other future research directions.

## 2 CONTEXT IN KERNEL DESIGN

Let  $\mathcal{S}_p = \{x_1^p, \dots, x_n^p\}$  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 kernel which, given two interest points  $(x_i^p, x_j^q)$ , provides a similarity measure between them. This will be designed as shown in Section (2.3). Our goal is to use  $k$  in order to build 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 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 subset kernel  $\mathcal{K}$ , also referred to as the convolution kernel, between  $\mathcal{S}_p = \{x_i^p\}_{i=1}^n$  and  $\mathcal{S}_q = \{x_j^q\}_{j=1}^m$  as  $\mathcal{K}(\mathcal{S}_p, \mathcal{S}_q) = \sum_i^n \sum_j^m k(x_i^p, x_j^q)$ .

Here  $k$  may be any symmetric and continuous function 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 [14]. Since  $\mathcal{K}$  is defined as the sum of all the cross similarities between all the possible sample pairs taken from  $\mathcal{S}_p \times \mathcal{S}_q$ , its evaluation does not require any (hard) alignment between these pairs. Nevertheless, the value of  $k(x_i^p, x_j^q)$  should ideally be high only if  $x_i^p$  actually matches  $x_j^q$  (see Fig. 1, rows 2, 4, right), so  $k$  needs to be appropriately designed while also guaranteeing the positive definiteness.

### 2.2 Context

Formally, an interest point  $x$  is defined as  $x = (\psi_g(x), \psi_f(x), \psi_o(x), \omega(x))$  where the symbol  $\psi_g(x) \in \mathbb{R}^2$  stands for the 2D coordinates of  $x$  while  $\psi_f(x) \in \mathbb{R}^s$  corresponds to the feature of  $x$  (for instance the 128 coefficients of the SIFT; [18]). 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 location, feature and orientation are considered different when they are not in the same image (since we want to take into account the context of the interest point in the image it belongs to).

Let  $d(x, x') = \|\psi_f(x) - \psi_f(x')\|_2$  measure the dissimilarity between two interest point features,  $\|\cdot\|_2$  is the

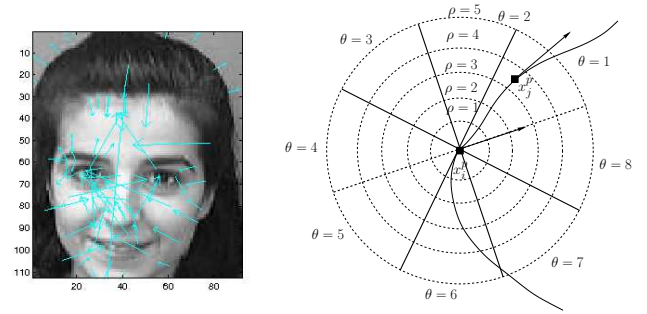


Fig. 2. This figure shows a collection of SIFT interest points (with their locations, orientations and scales) (left) and the partitioning of the context (also referred to as neighborhood) of an interest point into different sectors for orientations and bands for locations (right).

“entrywise”  $L_2$ -norm (i.e., the sum of the square values of vector coefficients). Introduce the context of  $x$

$$\mathcal{N}^{\theta, \rho}(x) = \{x' : \omega(x') = \omega(x), x' \neq x \text{ s.t. (1) and (2) hold}\},$$

with

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

and

$$\frac{\theta - 1}{N_a} \pi \leq \text{angle}(\psi_o(x), \psi_g(x') - \psi_g(x)) \leq \frac{\theta}{N_a} \pi. \quad (2)$$

Here  $\epsilon_p$  is the radius of a neighborhood disk surrounding  $x$  and  $\theta = 1, \dots, N_a$ ,  $\rho = 1, \dots, N_r$  correspond to indices of different parts of that disk (see Fig. 2). In practice,  $N_a$  and  $N_r$  correspond to 8 sectors and 8 bands. Notice that the definition of the neighborhood in this paper is different from the one proposed in [24], as the latter provides only a set of neighbors  $\mathcal{N}(x)$  around  $x$  which are not segmented into different parts. In [24],  $\mathcal{N}(x) = \cup_{\theta, \rho} \mathcal{N}^{\theta, \rho}(x)$ , and the new definition of neighborhoods  $\{\mathcal{N}^{\theta, \rho}(x)\}_{\theta, \rho}$  reflects the co-occurrence of different interest points with particular spatial geometric constraints (see again Fig. 2).

### 2.3 Context-Dependent 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  $x'$ :  $\mathbf{K}_{x, x'} = k(x, x')$ . 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  $g_{\theta, \rho}(x, x') = \mathbb{1}_{\{x' \in \mathcal{N}^{\theta, \rho}(x)\}}$ . Let  $\mathbf{D}_{x, x'} = d(x, x')$ .

We propose to use the kernel on  $\mathcal{X}$  defined by solving

$$\begin{aligned} \min_{\mathbf{K}} \quad & \text{Tr}(\mathbf{K} \mathbf{D}') + \beta \text{Tr}(\mathbf{K} \log \mathbf{K}') \\ & - \alpha \sum_{\theta, \rho} \text{Tr}(\mathbf{K} \mathbf{P}_{\theta, \rho} \mathbf{K}' \mathbf{P}'_{\theta, \rho}) \quad (3) \\ \text{s.t.} \quad & \begin{cases} \mathbf{K} \geq 0 \\ \|\mathbf{K}\|_1 = 1 \end{cases} \end{aligned}$$

Here  $\alpha, \beta \geq 0$  and the operations  $\log$  (natural) and  $\geq$  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  $\text{Tr}$  denotes matrix trace. The first term, in the above constrained minimization problem, measures the quality of matching two features  $\psi_f(x)$ ,  $\psi_f(x')$ . 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 interest points, the probability distribution  $\{k(x, x')\}$  should be flat so the negative of the entropy is minimized. This term also helps defining a direct analytic solution of the constrained minimization problem (3). 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 also makes it possible to consider the spatial configuration of the neighborhood of each interest point 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  $\mathcal{X} \times \mathcal{X}$ .

## 2.4 Solution

*Proposition 1:* 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

$$\zeta \exp(\zeta) < 1 \quad (4)$$

$$\|\exp(-\mathbf{D}/\beta)\|_1 \geq 2 \quad (5)$$

the optimization problem (3) 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

$$G(\mathbf{K}) = \exp \left\{ -\frac{\mathbf{D}}{\beta} + \frac{\alpha}{\beta} \sum_{\theta, \rho} (\mathbf{P}_{\theta, \rho} \mathbf{K} \mathbf{P}'_{\theta, \rho} + \mathbf{P}'_{\theta, \rho} \mathbf{K} \mathbf{P}_{\theta, \rho}) \right\}, \quad (6)$$

and

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

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

$$\|\mathbf{K}^{(t)} - \tilde{\mathbf{K}}\|_1 \leq L^t \|\mathbf{K}^{(0)} - \tilde{\mathbf{K}}\|_1. \quad (7)$$

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

By taking not too large  $\beta$ , one can ensure that (5) holds. Then by taking small enough  $\alpha$ , Inequality (4) 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  $\alpha = 0$ .

In comparison to [24], to have partitioned the neighborhood into several cells corresponding to different degrees of proximity (as shown in Fig. 2) has lead to significant improvements of our experimental results (see also Table. 3). On the one hand, the constraint (4) becomes easier to satisfy, for larger  $\alpha$  with partitioned neighborhood, compared to [24]. On the other hand, when the context is split into different parts, we end up with a context term (right-hand side term inside the exponential function in Eq. 6), which grows slowly compared to the one presented in our previous work [24] and grows only if *similar spatial configurations* of interest points have high kernel values. Therefore, numerically, the evaluation of that term is still tractable for large values of  $\alpha$  which apparently produces a more positively influencing (and precise) context-dependent term, i.e., last term in (3) (see also Table. 1, bottom and Fig. 4).

*Proof:*

Introduce the function defined on the set of matrices  $\mathbf{K}$  satisfying the constraints in (3)

$$\begin{aligned} F : \mathbf{K} \mapsto & \text{Tr}(\mathbf{K} \mathbf{D}') + \beta \text{Tr}(\mathbf{K} \log \mathbf{K}') \\ & - \alpha \sum_{\theta, \rho} \text{Tr}(\mathbf{K} \mathbf{P}_{\theta, \rho} \mathbf{K}' \mathbf{P}'_{\theta, \rho}). \end{aligned}$$

The function is defined by continuity at matrices  $\mathbf{K}$  for which there exists  $(x, x')$  such that  $\mathbf{K}_{x, x'} = 0$ . This is possible since  $\text{Tr}(\mathbf{K} \log \mathbf{K}') = \sum_{x, x'} \mathbf{K}_{x, x'} \log(\mathbf{K}_{x, x'})$  and since the function  $t \mapsto t \log t$  basically defined on the positive real numbers can be continuously extended at  $t = 0$  by setting  $0 \log(0) = 0$ .

Since the function  $t \mapsto t \log t$  has a derivative going to  $-\infty$  when  $t$  goes 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

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

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

$$G(\mathbf{K}) = \exp \left\{ -\frac{\mathbf{D}}{\beta} + \frac{\alpha}{\beta} \sum_{\theta, \rho} (\mathbf{P}_{\theta, \rho} \mathbf{K} \mathbf{P}'_{\theta, \rho} + \mathbf{P}'_{\theta, \rho} \mathbf{K} \mathbf{P}_{\theta, \rho}) \right\}, \quad (8)$$

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 (8).

*Lemma 1:* 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  $\psi : \mathcal{B} \rightarrow \mathcal{B}$  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 (8) admits a unique solution  $\tilde{\mathbf{K}}$ , and Inequality (7) holds.

*Proof of Lemma 1:* Let us start by the following property of  $L_1$ -projection on the  $L_1$ -sphere  $S = \{y : \|y\|_1 = 1\}$  in some Euclidean space: for any nonzero vector  $x$

$$\|x/\|x\|_1 - x\|_1 = \min_{y \in S} \|y - x\|_1.$$

This holds since the above minimum is reached for the vectors  $y$  in the intersection of  $S$  and an appropriate parallelepiped with axis parallel to the coordinate axis, and which admits  $x$  as a vertex and contains  $x/\|x\|_1$  (for instance, when  $\|x\|_1 \geq 1$ , it is the parallelepiped whose diagonal is  $[0; x]$ ).

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

$$\begin{aligned} & \|\psi(\mathbf{K}_2) - \psi(\mathbf{K}_1)\|_1 \\ &= \left\| \frac{\mathbf{G}_2}{\|\mathbf{G}_2\|_1} - \frac{\mathbf{G}_1}{\|\mathbf{G}_1\|_1} \right\|_1 \\ &\leq \left\| \frac{\mathbf{G}_2}{\|\mathbf{G}_2\|_1} - \frac{\mathbf{G}_2}{\|\mathbf{G}_1\|_1} \right\|_1 + \left\| \frac{\mathbf{G}_2}{\|\mathbf{G}_1\|_1} - \frac{\mathbf{G}_1}{\|\mathbf{G}_1\|_1} \right\|_1 \\ &= \min_{\mathbf{K} : \|\mathbf{K}\|_1 = 1} \left\| \mathbf{K} - \frac{\mathbf{G}_2}{\|\mathbf{G}_1\|_1} \right\|_1 + \left\| \frac{\mathbf{G}_2}{\|\mathbf{G}_1\|_1} - \frac{\mathbf{G}_1}{\|\mathbf{G}_1\|_1} \right\|_1 \\ &\leq \left\| \frac{\mathbf{G}_1}{\|\mathbf{G}_1\|_1} - \frac{\mathbf{G}_2}{\|\mathbf{G}_1\|_1} \right\|_1 + \left\| \frac{\mathbf{G}_2}{\|\mathbf{G}_1\|_1} - \frac{\mathbf{G}_1}{\|\mathbf{G}_1\|_1} \right\|_1 \\ &= \frac{2}{\|\mathbf{G}_1\|_1} \|\mathbf{G}_2 - \mathbf{G}_1\|_1 \\ &\leq \|\mathbf{G}_2 - \mathbf{G}_1\|_1, \end{aligned} \quad (9)$$

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  $\Delta G = |\mathbf{G}_2 - \mathbf{G}_1|$  and  $\Delta K = |\mathbf{K}_2 - \mathbf{K}_1|$  be the matrices defined by

$[\Delta G]_{x, x'} = |[G_2]_{x, x'} - [G_1]_{x, x'}|$  and  $[\Delta K]_{x, x'} = |[K_2]_{x, x'} - [K_1]_{x, x'}|$ . We have

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

Therefore we have

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

which implies

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

Now we trivially have

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

hence we obtain

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

Plugging this inequality into (9), we get

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

## 2.5 Positive Definiteness

A kernel  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is positive (semi-)definite 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 positive definite kernel guarantees the existence of a Reproducing Kernel Hilbert Space (RKHS) such that  $k(x, x') = \langle \phi(x), \phi(x') \rangle$ , where  $\phi$  is an explicit or implicit mapping function from  $\mathcal{X}$  to the RKHS, and  $\langle \cdot, \cdot \rangle$  is the dot kernel in the RKHS.

*Proposition 2:* The context-dependent kernels on  $\mathcal{X}$  defined in Proposition (1) 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 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  $\mathbf{A} = \frac{\alpha}{\beta} \sum_{\theta, \rho} (\mathbf{P}_{\theta, \rho} \mathbf{K} \mathbf{P}'_{\theta, \rho} + \mathbf{P}'_{\theta, \rho} \mathbf{K} \mathbf{P}_{\theta, \rho})$  is positive semi-definite. As a consequence, from [29, 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 matrix 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

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

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 taken as positive definite. 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 EXPERIMENTS

#### 3.1 Databases and Settings

In order to show the extra-value of CDK with respect to other kernels, 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 and the extremely challenging ImageClef@ICPR Photo Annotation database<sup>3</sup>. The latter contains 18,000 pictures split into 53 categories; a subset of 8,000 images was used for training and testing as ground truth is publicly available for this subset only. The Smithsonian database contains 35 leaf species, each one represented by 4 – 100 examples, resulting into 1,525 images while the Olivetti set 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. 3).

Interest points are extracted from all these databases and encoded using different features. For a matter of comparison, we tested SIFT [18], self similarity [30] and shape context [3]. SIFT descriptor contains 128 dimensions while both self similarity and shape context contain, in our case, 64 coefficients taken from 8 orientations and 8 scales<sup>4</sup>. All the local feature vectors (SIFT, self similarity and shape context) are normalized to 1.

Given a picture, the goal is to predict which concepts (object classes) are present into that picture. This task is commonly known as concept recognition. For this purpose, we trained “one-versus-all” SVM classifiers

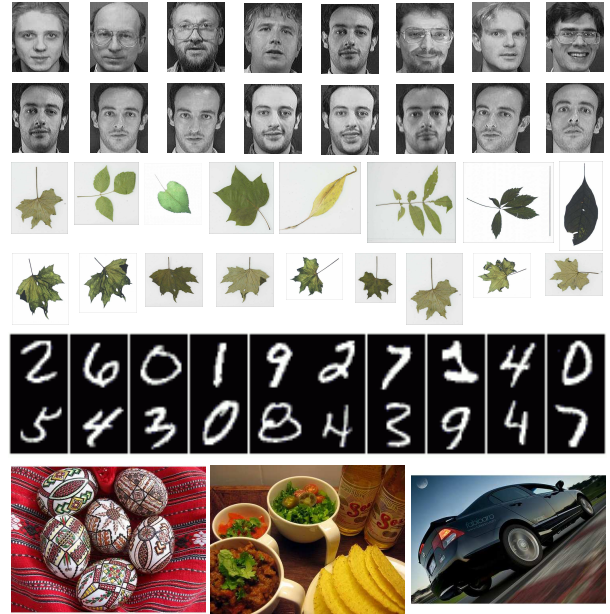


Fig. 3. This figure shows samples of training and test images taken respectively from the Olivetti face database, the Smithsonian leaf set, MNIST digit database and ImageClef@ICPR set.

for each concept; we use four random folds (80% of a database) for SVM training and the remaining fold for testing. We repeat this training process through different folds, for each concept, and we take the average error of the underlying SVM classifiers. This makes classification results less sensitive to sampling.

Performances are reported using the average hold out equal error rate (EER) on the test folds. EER is the balanced generalization error which equally weights errors in the positive and the negative sets, for a given concept. A smaller EER implies better performance. Note that this measure is evaluated, on the four databases, using a standard script provided by the ImageClef@ICPR evaluation campaign.

#### 3.2 Generalization and Comparison

We evaluate  $\mathcal{K}$  (see Section 2.1) and hence  $\mathbf{K}^{(t)}$ ,  $t \in \mathbb{N}^+$  using five power assist settings: (i) linear  $\mathbf{K}_{x, x'}^{(0)} = \langle \psi_f(x), \psi_f(x') \rangle$ , (ii) polynomial  $\mathbf{K}_{x, x'}^{(0)} = (\langle \psi_f(x), \psi_f(x') \rangle + 1)^2$ , (iii) Gaussian  $\mathbf{K}_{x, x'}^{(0)} = \exp(-d(x, x')/\beta)$ , (iv) Chi-square  $\mathbf{K}_{x, x'}^{(0)} = 1 - \frac{1}{2} \sum_i \frac{(\psi_f(x)_i - \psi_f(x')_i)^2}{(\psi_f(x)_i + \psi_f(x')_i)}$  (denoted Chi-Sq) and (v) Histogram intersection  $\mathbf{K}_{x, x'}^{(0)} = \sum_i \min(\psi_f(x)_i, \psi_f(x')_i)$  (denoted HI). 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-free kernels (i.e.,  $\mathbf{K}^{(t)}$ ,  $t = 0$ ). For this purpose, we trained the “one-versus-all” SVM classifiers for each class in Smithsonian, MNIST, Olivetti and ImageClef@ICPR sets using the subset kernel  $\mathcal{K}(\mathcal{S}_p, \mathcal{S}_q) = \sum_{x \in \mathcal{S}_p, x' \in \mathcal{S}_q} \mathbf{K}_{x, x'}^{(t)}$ . Again, performances are reported, on different test sets, using the hold-out equal error rate.

3. <http://www.imageclef.org/2010/ICPR/PhotoAnnotation/>

4. All these local descriptors were extracted using standard libraries in [www.robots.ox.ac.uk/~vgg/software/SelfSimilarity/](http://www.robots.ox.ac.uk/~vgg/software/SelfSimilarity/), [www.eecs.berkeley.edu/Research/Projects/CS/vision/shape/sc\\_digits.html](http://www.eecs.berkeley.edu/Research/Projects/CS/vision/shape/sc_digits.html) and [www.cs.ubc.ca/~lowe/keypoints/](http://www.cs.ubc.ca/~lowe/keypoints/)

Sets	Olivetti	MNIST	SmithSon	ImageClef
$\log_{10}\beta$	(EER+sd)	(EER+sd)	(EER+sd)	(EER+sd)
-2	1.56 ± 1.40	3.08 ± 1.52	3.59 ± 1.99	33.1 ± 8.35
-1	<b>1.37 ± 1.28</b>	<b>1.40 ± 1.24</b>	<b>3.39 ± 2.12</b>	<b>27.0 ± 8.64</b>
0	2.27 ± 1.75	15.2 ± 2.44	7.48 ± 2.99	36.6 ± 8.62
+1	28.7 ± 4.24	37.6 ± 3.07	25.7 ± 4.21	38.3 ± 8.85
Sets	Olivetti	MNIST	SmithSon	ImageClef
$\log_{10}\alpha$	(EER+sd)	(EER+sd)	(EER+sd)	(EER+sd)
-4	1.37 ± 1.28	1.40 ± 1.24	3.39 ± 2.12	27.0 ± 8.64
-3	1.37 ± 1.29	1.40 ± 1.24	3.39 ± 2.12	27.1 ± 8.64
-2	0.91 ± 0.33	1.17 ± 1.25	1.40 ± 0.88	25.4 ± 8.77
-1	<b>0.85 ± 0.10</b>	<b>1.17 ± 1.25</b>	<b>0.88 ± 0.80</b>	<b>23.3 ± 8.24</b>
0	1.03 ± 0.90	2.20 ± 2.46	1.85 ± 1.96	26.9 ± 8.59
	NC	NC	NC	NC

TABLE 1

The table in the top shows the error rate of the Gaussian term in CDK (i.e.,  $\exp(-\mathbf{D}/\beta)$ ) with respect to the scale parameter  $\beta$ . The table in bottom shows the error rate of CDK as a decreasing function of  $\alpha$ . Histogram intersection kernel is used for initialization and  $\beta = 0.1$ . Notice that when  $\log_{10}\alpha = -4$  (i.e.,  $\alpha \rightarrow 0$ ), results correspond to the context free kernel, and when  $\log_{10}\alpha = 0$  (i.e.,  $\alpha = 1$ ), the EER increases since the convergence of CDK to a fixed point is not guaranteed (NC stands for not convergent).

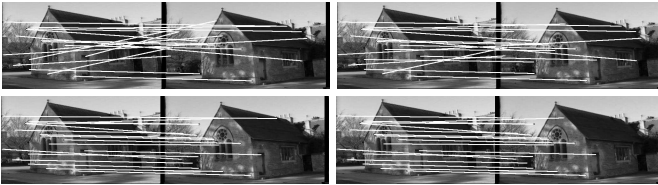


Fig. 4. This figure shows an example of the evolution of matching results for different and increasing values of  $\alpha$  (resp. from top-left to bottom-right, 0, 0.1, 2.5 and 3). We clearly see that when  $\alpha$  increases the matching results are better. We set  $\beta = 0.1$  and  $t = 1$  iteration only.

The setting of  $\beta$  is performed by maximizing the performance of the Gaussian term  $\exp(-\mathbf{D}/\beta)$  in CDK. For different databases, we found that the best performances are achieved for  $\beta = 0.1$  (see Table. 1, top) and this also guarantees condition (5) in practice. The influence (and the performance) of the context term in  $\mathbf{K}^{(t)}$  increases as  $\alpha$  increases (see Table. 1, bottom and Fig. 4), nevertheless and as shown earlier, the convergence of  $\mathbf{K}^{(t)}$  to a fixed point is guaranteed only if Eq. (4) is satisfied. Intuitively, the parameter  $\alpha$  should then be relatively high while also satisfying condition (4). Larger values of  $\alpha$  do not guarantee convergence of CDK and the classification performance may not converge to the best ones (see Table. 1, bottom, last row).

Table. (2), shows EERs of different combinations of local features including SIFT, self similarity (denoted "SSim") and shape context (denoted "SCont") with different context-free kernels (Linear, Gaussian, etc.) and

$\mathbf{K}^{(0)}$	Linear	Poly	Gauss	Chi-Sq	HI
Databases + Local Features +Kernels	(EER)	(EER)	(EER)	(EER)	(EER)
<b>MNIST</b>					
SIFT [18] + CFK	5.20%	5.15%	4.58%	3.96%	2.75%
SIFT [18] + CDK	4.97%	4.68%	4.15%	2.98%	2.69%
SSim [30] + CFK	4.91%	2.42%	2.69%	1.40%	1.40%
SSim [30] + CDK	3.86%	1.87%	1.64%	1.17%	1.17%
SCont [3] + CFK	8.77%	8.65%	7.60%	4.80%	4.74%
SCont [3] + CDK	3.77%	2.63%	2.57%	2.81%	2.34%
<b>Olivetti</b>					
SIFT [18] + CFK	1.70%	1.83%	1.63%	1.39%	1.37%
SIFT [18] + CDK	0.85%	0.85%	0.85%	0.85%	0.85%
SSim [30] + CFK	2.28%	2.48%	2.28%	1.72%	1.68%
SSim [30] + CDK	0.88%	0.85%	0.85%	0.85%	0.85%
SCont [3] + CFK	6.00%	2.52%	2.46%	1.65%	1.73%
SCont [3] + CDK	2.05%	1.20%	1.15%	1.09%	1.67%
<b>Smithsonian</b>					
SIFT [18] + CFK	9.96%	5.63%	4.31%	3.08%	3.39%
SIFT [18] + CDK	1.16%	1.06%	1.01%	0.88%	0.88%
SSim [30] + CFK	14.8%	12.8%	7.67%	3.33%	3.96%
SSim [30] + CDK	5.20%	5.10%	4.09%	1.67%	2.64%
SCont [3] + CFK	8.76%	3.21%	3.11%	6.28%	2.51%
SCont [3] + CDK	3.19%	1.73%	1.08%	1.03%	1.70%
<b>ImageClef</b>					
SIFT [18] + CFK	33.2%	31.1%	30.2%	27.2%	27.0%
SIFT [18] + CDK	28.3%	25.9%	25.4%	24.2%	23.3%
SSim [30] + CFK	32.1%	31.2%	30.6%	29.5%	29.1%
SSim [30] + CDK	27.8%	27.9%	27.7%	27.0%	25.2%
SCont [3] + CFK	34.4%	34.2%	33.2%	33.2%	33.4%
SCont [3] + CDK	34.2%	33.6%	33.0%	30.4%	30.5%

TABLE 2

This figure shows the EER on MNIST, Smithsonian, Olivetti and ImageClef@ICPR sets, using different local features and kernel settings. SSim, SCont stand respectively for self similarity and shape context local features. In all these experiments, we set  $\beta = \alpha = 0.1$ .

the underlying context-dependent ones. These experiments are shown for different databases and they clearly and consistently illustrate the out-performance of CDKs with respect to CFKs for almost all the features and the test sets, with only few iterations ( $t \leq 3$  in practice). Fig. 5 shows these EER class by class on the ImageClef@ICPR dataset; in almost all the classes, CDK decreases errors. Table. 3 shows also a comparison of our CDK with other kernels including baseline ones (Gaussian, histogram intersection, etc.), the kernel in [7], pyramid match kernel<sup>5</sup> [12] and our previous version of CDK [24].

## 4 DISCUSSION

**Invariance.** The adjacency matrices  $\mathbf{P}_{\theta,\rho}$  in  $\mathbf{K}^{(t)}$ , provide the intrinsic properties and also characterize 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 Eq. 1) is adapted to the scales of  $\psi_g(\mathcal{S}_p)$ . It follows that the context term of our kernel is invariant to any 2D similarity transformation. Notice, also, that the Gaussian term of  $\mathbf{K}^{(t)}$  may involve similarity invariant features  $\psi_f(\cdot)$  (such as SIFT features), so  $\mathbf{K}^{(t)}$  is similarity invariant.

5. <http://people.csail.mit.edu/jjl/libpmk/>

Database Method	Olivetti (EER)	MNIST (EER)	SmithSon (EER)	ImageClef (EER)
Baselines (CFK)				
Gauss	1.63%	2.69%	4.31%	30.2%
Chi-Sq	1.39%	1.40%	3.08%	27.2%
HI	1.37%	1.40%	3.39%	27.0%
CDK in [24]				
$\mathbf{K}^{(0)} \leftarrow$ Gauss	1.17%	2.75%	1.16%	28.0%
$\mathbf{K}^{(0)} \leftarrow$ Chi-Sq	1.09%	1.35%	1.18%	27.1%
$\mathbf{K}^{(0)} \leftarrow$ HI	1.04%	1.26%	0.90%	25.3%
CDK (current)				
$\mathbf{K}^{(0)} \leftarrow$ Gauss	<b>0.85%</b>	1.64%	1.01%	25.4%
$\mathbf{K}^{(0)} \leftarrow$ Chi-Sq	<b>0.85%</b>	<b>1.17%</b>	<b>0.88%</b>	24.2%
$\mathbf{K}^{(0)} \leftarrow$ HI	<b>0.85%</b>	<b>1.17%</b>	<b>0.88%</b>	<b>23.3%</b>
Local Kernel of [7]	1.66%	1.46%	2.96%	27.3%
Pyramid Match of [12]	1.20%	1.29%	2.78%	25.8%

TABLE 3

This table shows a comparison of the current CDK with the previous one, in [24] (using unpartitioned context), and also with respect to the local kernel introduced in [7] and the pyramid match kernel [12]. For all the CDKs,  $\alpha$  and  $\beta$  are set to 0.1.

**Performance.** 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, histogram intersection and Chi-square kernels, and few iterations ( $\leq 3$ ) for the other kernels (linear and polynomial). On the one hand, this corroborates the fact that the Gaussian, histogram intersection and Chi-square kernels provide state of the art performances, and on the other hand, their performances can be consistently improved by including the geometry and the context of objects.

**Runtime.** One limitation of our previous CDK [24] resides in its evaluation complexity. Assuming  $\mathbf{K}^{(t-1)}$  known for a given pair  $x, x'$ , the worst complexity is  $O(\max(N^2, s))$ , 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  $N < s^{\frac{1}{2}}$ , the complexity of evaluating CDK is strictly equivalent to that of usual kernels such as the linear. Nevertheless, the worst case ( $N \gg s^{\frac{1}{2}}$ ) makes CDK evaluation prohibitive and this is mainly due to the context term of  $\mathbf{K}_{x,x'}^{(t)}$ . A simple preprocessing step of feature clustering makes it possible to replace each feature by its closest cluster center among  $C$  possible candidates (Here  $C$  is fixed and small while  $N$  varies and large). Instead of summing  $N^2$  terms in CDK, we sum only  $C^2$  terms ( $C \ll N$ ) and this makes CDK training much faster. In practice, it takes about 5 hours to evaluate CDK on the 8,000 pictures from the ImageClef@ICPR set, instead of three days without clustering. This feature clustering does not only speed up CDK evaluation but also makes features "coarse" and more suitable for generic databases.

**Relation to diffusion maps.** One may show that CDK

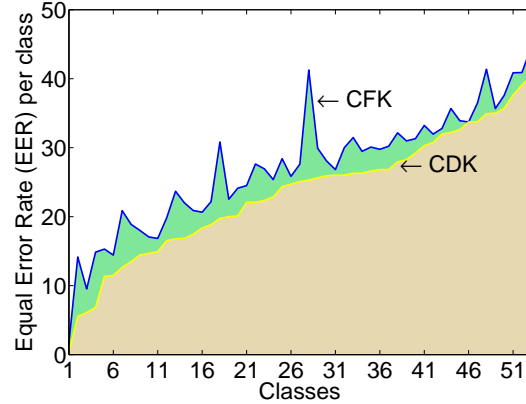


Fig. 5. This figure shows the EER of CDK class by class on the ImageClef@ICPR dataset ( $\alpha = \beta = 0.1$  and histogram intersection is used for initialization). The results of the other participants (excluding our CDK results) may be found in <http://www.imageclef.org/2010/ICPR/PhotoAnnotationEERResults>. In the x-labels, classes were sorted according to the underlying CDK EER (from the simplest to the hardest one). This order corresponds to "Desert", "River", "Sea", "Beach-Holidays", "Mountains", "Sunset-Sunrise", "Snow", "Lake", "Motion-Blur", "Landscape-Nature", "Water", "Clouds", "Autumn", "Small-Group", "Partylife", "Flowers", "Overexposed", "Portrait", "Sky", "Out-of-focus", "Night", "Winter", "Family-Friends", "Macro", "Trees", "Partly-Blurred", "Food", "Spring", "Still-Life", "Citylife", "Outdoor", "Animals", "Single-Person", "Underexposed", "Fancy", "Big-Group", "Building-Sights", "No-Blur", "Indoor", "No-Visual-Time", "Plants", "Day", "No-Visual-Place", "Canvas", "No-Visual-Season", "Sports", "No-Persons", "Neutral-Illumination", "Sunny", "Summer", "Vehicle", "Aesthetic-Impression" and "Overall-Quality".

captures topology of context through diffusion maps [17]. Without loss of generality, let's assume  $\mathbf{P}_{\theta,\rho}$  independent of  $\theta, \rho$  and let's denote it simply as  $\mathbf{P}$ . Considering  $\beta \gg 2\alpha$  and using the first order Taylor expansion one may approximate the kernel  $\mathbf{K}^{(t)}$  by

$$-\sum_{k=0}^{t-1} \frac{1}{\beta} \left(\frac{2\alpha}{\beta}\right)^k \mathbf{P}^{(k)} \mathbf{D} \mathbf{P}^{(k)'} + \left(\frac{2\alpha}{\beta}\right)^t \mathbf{P}^{(t)} \mathbf{K}^{(0)} \mathbf{P}^{(t)'}. \quad (10)$$

Let  $\mathcal{G}$  be the graph defined by the adjacency matrix  $\mathbf{P}$  and let the entry  $\mathbf{P}_{ij} = P_1(j|i)$  denote the probability of a 1-step random walk from a node (an interest point)  $x_i^p$  to  $x_j^p$  in  $\mathcal{G}$ . The idea behind diffusion maps [17], is to represent higher order walks by taking powers of  $\mathbf{P}$ , i.e.,  $\mathbf{P}^{(k)} = \mathbf{P}^{(k-1)} \mathbf{P}$ . Here  $\mathbf{P}^{(k)}$  is the k-step random walk transition matrix which models a Markovian process; the k-step transition likelihood is the sum over all the possible k-1 steps linking  $x_i^p$  to  $x_j^p$  ( $\mathbf{P}_{ij}^{(k)} = P_k(j|i) = \sum_{\ell=1}^n P_{k-1}(\ell|i) P_1(j|\ell)$ ). In this definition,  $k$  acts as a scale factor that increases the local influence of the context when designing CDK. For a given  $t$ , the right-hand side term of (10) is the "t-step" similarity between contexts of interest points, inside a manifold (with a topology

defined by diffusion maps [17]) while the left-hand side term takes into account the visual similarity between interest points in their feature spaces (such as the SIFT). Put differently, when CDK converges, it models visual similarity of interest points and topology of their context.

## 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 positive definite and applicable to SVM learning. This is achieved for object classification problems and has better performance compared to SVMs with context-free kernels.

The proposed approach, even though presented for kernel design, might be straightforwardly extended to graph matching. Indeed, one may define graph adjacency matrices and use exactly the same energy as (3) in order to derive similarity between nodes belonging to two different graphs. Obviously, matches correspond to pairs which maximize the CDK values. The approach may also be extended to other pattern analysis problems such as bioinformatics, speech, text, machine translation, etc.

## ACKNOWLEDGEMENTS

This work was supported in part by a grant from the Research Agency ANR (Agence Nationale de la Recherche) under the AVEIR and the MGA Projects.

## REFERENCES

- [1] C. Bahlmann, B. Haasdonk, and H. Burkhardt. On-line handwriting recognition with support vector machines, a kernel approach. *IWFHR*, pages 49–54, 2002.
- [2] 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.
- [3] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. *NIPS*, 2000.
- [4] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. In *CVPR*, 1:26–33, 2005.
- [5] 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.
- [6] S. Boughorbel. *Kernels for Image Classification with Support Vector Machines*. PhD thesis, Faculte d’Orsay, 2005.
- [7] B. Caputo, C. Wallraven, and M.-E. Nilsback. Object categorization via local kernels. *ICPR*, 2:132–135, 2004.
- [8] O. Chapelle, P. Haffner, and V. Vapnik. Svms for histogram-based image classification. *Transaction on Neural Networks*, 10(5), 1999.
- [9] M. Cuturi. Etude de noyaux de semigroupe pour objets structures dans le cadre de l’apprentissage statistique. *PhD thesis Gostatistique*, ENSMP, 2005.
- [10] T. Gartner. A survey of kernels for structured data. *Multi Relational Data Mining*, 5(1):49–58, 2003.
- [11] G. Genton. Classes of kernels for machine learning: A statistics perspective. *JMLR*, 2(12):299–312, 2001.
- [12] 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.
- [13] C. Harris and M. Stephens. A combined corner and edge detector. *Alvey Vision Conference*, pages 147–151, 1988.
- [14] D. Haussler. Convolution kernels on discrete structures. *Technical Report UCSC-CRL-99-10*, University of California in Santa Cruz, Computer Science Department, July, 1999.
- [15] T. Jaakkola, M. Diekhans, and D. Haussler. Using the fisher kernel method to detect remote protein homologies. *ISMB*, pages 149–158, 1999.
- [16] R. Kondor and T. Jebara. A kernel between sets of vectors. In *proceedings of the 20th ICML conference*, 2003.
- [17] S. Lafon, Y. Keller, and R. R. Coifman. Data fusion and multi-cue data matching by diffusion maps. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(11):1784–1797, 2006.
- [18] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [19] S. Lyu. Mercer kernels for object recognition with local features. In *the proceedings of the IEEE CVPR conference*, 2005.
- [20] 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.
- [21] P. Moreno, P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *Neural Information Processing Systems*, 2003.
- [22] 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.
- [23] J.-Y. Nie, M. Simard, P. Isabelle, and R. 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.
- [24] H. Sahbi, J. Audibert, J. Rabarisoa, and R. Keriven. Context dependent kernel design for object matching and recognition. in *IEEE CVPR Conference*, 2008.
- [25] C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.
- [26] B. Scholkopf and A. Smola. *Learning with kernels: Support vector machines, regularization, optimization and beyond*. MIT Press, Cambridge, MA, 2002.
- [27] B. Scholkopf, K. Tsuda, and J.-P. Vert. *Kernel methods in computational biology*. MIT Press, 2004.
- [28] A. Shashua and T. Hazan. Algebraic set kernels with application to inference over local image representations. In *Neural Information Processing Systems (NIPS)*, 2004.
- [29] J. Shawe-Taylor and N. Cristianini. *Support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.
- [30] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition 2007 (CVPR’07)*, June 2007.
- [31] K. 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.
- [32] M. Swain and D. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.
- [33] 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.
- [34] V. N. Vapnik. *Statistical learning theory*. A Wiley-Interscience Publication, 1998.
- [35] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. *ICCV*, pages 257–264, 2003.
- [36] L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *Journal of Machine Learning Research*, 4:913–931, 2003.



**Hichem Sahbi** Hichem Sahbi received his MSc degree in theoretical computer science from the University of Paris Sud, in Orsay, France in 1999, and his PhD in computer vision and machine learning from INRIA/Versailles University, France in 2003. From 2003-2006 he was a research associate first at the Fraunhofer Institute in Darmstadt, Germany, and then at the Machine Intelligence Laboratory at Cambridge University, UK. From 2006-2007, he was a senior research associate at l'Ecole des Ponts ParisTech, in

Paris. Since 2007, he has been a CNRS CR1 associate professor at Télécom ParisTech/ENST, in Paris. His research interests include statistical machine learning, kernel and graph based inference, computer vision, and image retrieval.



**Jean-Yves Audibert** Jean-Yves Audibert received the PhD degree in mathematics from the University of Paris 6 in 2004. Since then, he is a researcher in the Computer Science department at Ecole des Ponts ParisTech. Since 2007, he is also a research associate in the Computer Science department at Ecole Normale Supérieure in a joint INRIA/ENS/CNRS project. His research interest and publications range from Statistics to Computer Vision, including theoretical properties of learning procedures, boosting algorithms,

kernel machines, object recognition, image segmentation, content-based image retrieval.



**Renaud Keriven** Renaud Keriven is professor of Computer Science at Ecole des Ponts ParisTech where he heads the IMAGINE group. He is associate professor at Ecole Polytechnique. He graduated from Ecole Polytechnique Paris in 1988 and received the PhD degree from Ecole des Ponts ParisTech in 1997. From 2002 to 2007, he was assistant director of the Odysse team (leader Pr. O. Faugeras) at Ecole Normale Supérieure, Paris. His research interests include Multiview Stereovision, Shape priors, Level Set

and Graph Cut Methods, Discrete and continuous optimizations and Generic Programming on Graphics Processing Units.