Context-dependent kernel design for object matching and recognition

Hichem Sahbi, Jean-Yves Audibert, Jaonary Rabarisoa, Renaud Keriven

Context-Dependent Kernel Design for Object Matching and Recognition

Hichem Sahbi
UMR 5141, CNRS
Telecom ParisTech, France
sahbi@telecom-paristech.fr

Jean-Yves Audibert, Jaonary Rabarisoa and Renaud Keriven
Certis Lab
ENPC ParisTech, France
{audibert,rabariso,keriven}@certis.enpc.fr

Abstract

The success of kernel methods including support vector networks (SVMs) strongly depends on the design of appropriate kernels. While initially kernels were designed in order to handle fixed-length data, their extension to unordered, variable-length data became more than necessary for real pattern recognition problems such as object recognition and bioinformatics.

We focus in this paper on object recognition using a new type of kernel referred to as "context-dependent". Objects, seen as constellations of local features (interest points, regions, etc.), are matched by minimizing an energy function mixing (1) a fidelity term which measures the quality of feature matching, (2) a neighborhood criteria 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 also satisfies the Mercer condition. Experiments conducted on object recognition show that when plugging our kernel in SVMs, we clearly outperform SVMs with "context-free" kernels.

1. Introduction

Object recognition is one of the biggest challenges in vision and its interest is still growing [10]. Among existing methods, those based on machine learning (ML), show a particular interest as they are performant and theoretically well grounded [5]. ML approaches, such as the popular support vector networks [6], basically require the design of similarity measures, also referred to as *kernels*, which should provide high values when two objects share similar structures/appearances and should be invariant, as much as possible, to the linear and non-linear transformations. Kernel-based object recognition methods were initially *holistic*, i.e., each object is mapped into one or multiple fixed-length vectors and a similarity, based on color, texture or shape [29, 8], is then defined. *Local* kernels, i.e., those based on bags or local sets were introduced in order to represent data which cannot be represented by ordered and fixed-length feature vectors, such as graphs, trees, interest points, etc [11]. 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[1]. While the design of kernels gathering flexibility, invariance and low complexity is a challenging task; the proof of their positive definiteness is sometimes harder [9]. This property also known as the Mercer condition ensures, according to Vapnik's SVM theory [30], optimal generalization performance and also the uniqueness of the SVM solution.

Consider a database of objects (images), each one seen as a constellation of local features, for instance interest points [24, 19, 18], extracted using any suitable filter [13]. 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 [3]. 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 [17, 21] 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 [21] and the "Bhattacharyya affinity" in [17]. Only the function in [17] satisfies the Mercer condition and both kernels were applied for image recognition tasks. In [33], 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

[1] In case of object recognition, invariance means robustness to occlusion, geometric transformations and illumination.

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 [26]. 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 [32] considers the similarity function, between two feature sets, as the sum of their matching scores and unlike discussed in [32] this kernel is actually not Mercer [2]. In [20], 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 [7], 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, [12] 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 [2], an effective matching strategy for handwritten character recognition, nevertheless the Mercer condition is not guaranteed.

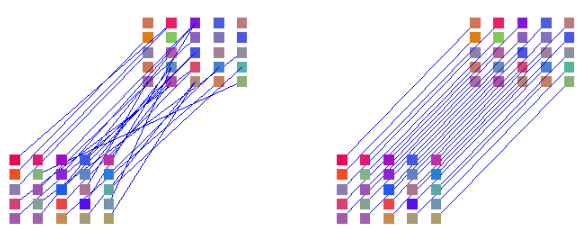

Figure 1. This figure shows a comparison of the matching results when using a naive matching strategy without geometry, (which consists in finding the set of possible matches by minimizing a distance between the color descriptors) and our "context-dependent" matching.

Naive matching	**'H'**	**'i'**		**'S'**	**'i'**	**'r'**
'S'	0	0	-	1	0	0
'i'	0	1	-	0	1	0
'r'	0	0	-	0	0	1
Context-dependent	-	-	-	-	-	-
'S'	0	0	-	.38	0	0
'i'	0	.36	-	0	.39	0
'r'	0	0	-	0	0	.38

Table 1. This table shows a simple comparison between similarity measures when using naive matching (upper table) and context-dependent matching (lower table).

1.1. Motivation and Contribution

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 (see Figures 1 and 2, left). The same argument is supported in [24], for the general problem of visual features matching, about the strong spatial correlation 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 Table 1). *Hence, it is necessary to consider the context as a part of the alignment process when designing kernels.*

In this paper, we introduce a new kernel, called "context-dependent" (or "CDK") and defined as the fixed-point of an energy function which balances an "alignment quality" term and a "neighborhood" criteria. The alignment quality is inversely proportional to the expectation of the Euclidean distance between the most likely aligned features (see Section 2) while the neighborhood criteria measures the spatial coherence of the alignments; given a pair of features (f_p, f_q) with a high alignment quality, the neighborhood criteria is proportional to the alignment quality of all the pairs close[2] to (f_p, f_q). *The general form of "CDK" captures the similarity between any two features by incorporating also their context, i.e., the similarity of the surrounding features.* Our proposed kernel can be

[2]The closeness is defined in Section 2.

viewed as a variant of "dynamic programming" kernel [2] where instead of using the ordering assumption we consider a neighborhood assumption which states that two points match if they have similar features and if they satisfies a neighborhood criteria i.e., their neighbors 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" 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 context in the matching process. Notice that the goal of this paper is not to extend local features to be global and doing so (as in [22, 1]) makes local features less invariant, but rather to design a similarity kernel ("CDK") which captures the context while being invariant. Even though we investigate "CDK" 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 documents retrieval [23], machine translation [28] and bioinformatics [25].

In the remainder of this paper we consider the following terminology and notation. A feature refers to a local interest point $x_i^p = (\psi_g(x_i^p), \psi_f(x_i^p), y_p)$, here i stands for the i^{th} sample of the subset $\mathcal{S}_p = \{x_1^p, \dots, x_n^p\}$ and $y_p \in \mathbb{N}^+$ is a unique indicator which provides the class or the subset including x_i^p. $\psi_g(x_i^p) \in \mathbb{R}^2$ stands for the $2D$ coordinates of the interest-point x_i^p while $\psi_f(x_i^p) \in \mathbb{R}^s$ corresponds to the descriptor of x_i^p (for instance the 128 coefficients of the SIFT[19]). We define $\mathcal{X}$ as the set of all possible features taken from all the possible images in the world and X is a random variable standing for a sample in $\mathcal{X}$. We also consider $k_t : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ as a symmetric function which, given two samples (x_i^p, x_j^q), provides a similarity measure. Other notations will be introduced as we go along through different sections of this paper which is organized as follows. We first introduce in Section 2, our energy function which makes it possible to design our context-dependent kernel 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 string matching and machine translation. We conclude in Section 5 and we provide some future research directions.

2. Kernel Design

Define $\mathcal{X} = \cup_{p \in \mathbb{N}^+} \mathcal{S}_p$ as the set of all possible interest points taken from all the possible objects in the world. We assume that all the objects are sampled with a given cardinality i.e., $|\mathcal{S}_p| = n$, $|\mathcal{S}_q| = m$, $\forall\, p, q \in \mathbb{N}^+$ (n and m might be different). Our goal is to design a kernel K which provides the similarity between any two objects (subsets) $\mathcal{S}_p$, $\mathcal{S}_q$ in $\mathcal{X}$.

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 K between $\mathcal{S}_p = \{x_1^p, \dots, x_i^p, \dots, x_n^p\}$ and $\mathcal{S}_q = \{x_1^q, \dots, x_j^q, \dots, x_m^q\}$ as*

$$K(\mathcal{S}_p, \mathcal{S}_q) = \sum_i^n \sum_j^m k\left(x_i^p, x_j^q\right), \tag{1}$$

here k is symmetric and continuous on $\mathcal{X} \times \mathcal{X}$, so K will also be continuous and symmetric. Since 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 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 $K(\mathcal{S}_p, \mathcal{S}_q)$ should be dominated by $\max_{i,j} k\left(x_i^p, x_j^q\right)$, so k should be appropriately designed (see Section 2.1).

Let X be a random variable standing for samples taken from $\mathcal{S}_p$ and X' is defined in a similar way for the subset $\mathcal{S}_q$. We design our kernel $k(x_i^p, x_j^q) = \mathbb{P}(X' = x_j^q, X = x_i^p)$ as the *joint probability* that x_j^q matches x_i^p. Again, it is clear enough (see Figures 1,2 and Table 1) that when this joint probability is estimated using only the sample coordinates (without their contexts), this may result into many false matches and wrong estimate of $\left\{\mathbb{P}(X' = x_j^q, X = x_i^p)\right\}_{i,j}$.

Before describing the whole design of k, we start with our definition of context-dependent kernels.

Definition 2 (Context-Dependent Kernels) *we define a context-dependent kernel k as any symmetric, continuous and recursive function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ such that $k(x_i^p, x_j^q)$ is equal to*

$$c(x_i^p, x_j^q) \times h\left(\sum_{k,\ell} k(x_k^p, x_\ell^q)\ \mathbb{V}\left(x_i^p, x_k^p, x_j^q, x_\ell^q\right)\right), \tag{2}$$

here c is a positive (semi) definite and context-free (non-recursive) kernel, $\mathbb{V}(x, x', y, y')$ is a monotonic decreasing function of any (pseudo) distance involving (x, x', y, y') and $h(x)$ is monotonically increasing.

2.1. Approach

We consider the issue of designing k using a variational framework. Let $\mathcal{I}_p = \{1, \dots, n\}$, $\mathcal{I}_q = \{1, \dots, m\}$, $\mu = \{k(x_i^p, x_j^q)\}$, $d(x_i^p, x_j^q) = \|\psi_f(x_i^p) - \psi_f(x_j^q)\|_2$ and $\mathcal{N}_p(x_i^p) = \{x_k^p \in \mathcal{S}_p : k \neq i,\ \|\psi_g(x_i^p) - \psi_g(x_k^p)\|_2 \leq \epsilon_p\}$

(ϵ_p defines a neighborhood and $\mathcal{N}_q$ is defined in the same way for $\mathcal{S}_q$). Consider $\alpha, \beta \geq 0$, $\mu = \{k(x_i^p, x_j^q)\}$ is found by solving

$$
\begin{aligned}
\min_{\mu} \quad & \sum_{i \in \mathcal{I}_p, j \in \mathcal{I}_q} k(x_i^p, x_j^q)\, d(x_i^p, x_j^q) + \\
& \beta \sum_{i \in \mathcal{I}_p, j \in \mathcal{I}_q} k(x_i^p, x_j^q)\, \log(k(x_i^p, x_j^q)) + \\
& \alpha \sum_{i \in \mathcal{I}_p, j \in \mathcal{I}_q} k(x_i^p, x_j^q) \left(- \sum_{\substack{x_k^p \in \mathcal{N}_p(x_i^p),\\ x_\ell^q \in \mathcal{N}_q(x_j^q)}} k(x_k^p, x_\ell^q) \right) \\
\text{s.t.} \quad & k(x_i^p, x_j^q) \in [0,1] \quad i \in \mathcal{I}_p, \quad j \in \mathcal{I}_q \\
& \sum_{i,j} k(x_i^p, x_j^q) = 1
\end{aligned}
\tag{3}
$$

The first term measures the quality of matching two descriptors $\psi_f(x_i^p)$, $\psi_f(x_j^q)$. In the case of SIFT, this is considered as the distance, $d(x_i^p, x_j^q)$, between the 128 SIFT coefficients of x_i^p and x_j^q. A high value of $d(x_i^p, x_j^q)$ should result into a small value of $k(x_i^p, x_j^q)$ and vice-versa. The second term is a regularization criteria which considers that without any a priori about the aligned samples, the probability distribution $\{k(x_i^p, x_j^q)\}$ 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 (see. appendix). The third term is a neighborhood criteria which considers that a high value of $k(x_i^p, x_j^q)$ should imply high kernel values in the neighborhoods $\mathcal{N}_p(x_i^p)$ and $\mathcal{N}_q(x_j^q)$. This criteria makes it possible to consider the context (spatial configuration) of each sample in the matching process.
We formulate the minimization problem by adding an equality constraint and bounds which ensure that $\{k(x_i^p, x_j^q)\}$ is a probability distribution.

Proposition 1 *(3) admits a solution in the form of a context-dependent kernel* $k_t(x_i^p, x_j^q) = v_t(x_i^p, x_j^q)/Z_t$, *with* $t \in \mathbb{N}^+$, $Z_t = \sum_{i,j} v_t(x_i^p, x_j^q)$ *and* $v_t(x_i^p, x_j^q)$ *defined as*

$$
\begin{aligned}
& \exp\left(-\frac{d(x_i^p, x_j^q)}{\beta} - 1\right) \times \\
& \quad \exp\left(\frac{2\alpha}{\beta} \sum_{k,\ell} \mathbb{V}(x_i^p, x_k^p, x_j^q, x_\ell^q)\ \ k_{t-1}(x_k^p, x_\ell^q)\right)
\end{aligned}
\tag{4}
$$

which is also a Gibbs distribution.

Proof. see appendix.
In (4), we set v_0 to any positive definite kernel (see proposition 3) and we define $\mathbb{V}(x_i^p, x_k^r, x_j^q, x_\ell^s)$ as $g(x_i^p, x_k^r) \times g(x_j^q, x_\ell^s)$ where g is a decreasing function of any (pseudo) distance involving (x_i^p, x_k^r), *not necessarily symmetric*. In practice, we consider $g(x_i^p, x_k^r) = \mathbb{1}_{\{r=p\}} \times \mathbb{1}_{\{x_k^r \in \mathcal{N}_p(x_i^p)\}}$.

It is easy to see that k_t is a P-kernel on any $\mathcal{S}_p \times \mathcal{S}_q$ [14] (as the joint probability over sample pairs taken from any $\mathcal{S}_p$ and $\mathcal{S}_q$ sums to one), so the value of the subset kernel $K(\mathcal{S}_p, \mathcal{S}_q)$ defined in (1) is constant and *useless*. To make k_t (up to a factor) a P-kernel on $\mathcal{X} \times \mathcal{X}$ (and not on $\mathcal{S}_p \times \mathcal{S}_q$), we cancel the equality constraint in (3) and we can prove in a similar way (see. appendix) that $k_t(x_i^p, x_j^q)$ is equal to $v_t(x_i^p, x_j^q)$ which is still a context-dependent kernel.

2.2. Mercer Condition

Before stating our result about the positive definiteness of k_t and also K, we remind some elementary definitions and results. Let $\mathcal{X}$ be an input space and let $k_t : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be symmetric and continuous. k_t is Mercer, i.e., positive (semi) definite, if and only if any Gram (kernel scalar product) matrix built by restricting k_t to any finite subset of $\mathcal{X}$ is positive (semi) definite. A Mercer kernel k_t guarantees the existence of a reproducing kernel Hilbert space $\mathcal{H}$ where k_t can be written as a dot product i.e., $\exists \Phi_t : \mathcal{X} \rightarrow \mathcal{H}$ such that $\forall\, x, x' \in \mathcal{X}$, $k_t(x, x') = \langle \Phi_t(x), \Phi_t(x') \rangle$.

Proposition 2 (Closure[27]) *the sum and the product of any two Mercer kernels is a Mercer kernel. The exponential of any Mercer kernel is also a Mercer kernel.*

Proof. see, for instance, [27].
Now, let us state *our result* about the positive definiteness of the "CDK" kernel.

Proposition 3 *let* $\mathbb{V}(x_i^p, x_k^p, x_j^q, x_\ell^q) = g(x_i^p, x_k^p) g(x_j^q, x_\ell^q)$, *consider* $g : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ *and* k_0 *positive definite. The kernel* k_t *is then positive definite.*

Proof. initially ($t = 0$), k_0 is per definition a positive definite kernel. By induction, let us assume k_{t-1} a Mercer kernel i.e., $\exists \Phi_{t-1} : k_{t-1}(x, x') = \langle \Phi_{t-1}(x), \Phi_{t-1}(x') \rangle$, $\forall x, x' \in \mathcal{X}$. Now, the sufficient condition will be to show that $\left(\sum_{y,y'} \mathbb{V}(x, y, x', y')\, k_{t-1}(y, y')\right)$ is also a Mercer kernel. Then, by the closure of the exponential and the product (see proposition 2), k_t will then be Mercer.
We need to show

$$
\begin{aligned}
& \forall x_1, \ldots, x_d \in \mathcal{X}, \ \ \forall c_1, \ldots, c_d \in \mathbb{R}, \\
& (*) = \sum_{i,j} c_i\, c_j \left(\sum_{y,y'} \mathbb{V}(x_i, y, x_j, y')\, k_{t-1}(y, y') \right) \geq 0
\end{aligned}
\tag{5}
$$

We have

$$
\begin{aligned}
(*) &= \sum_{i,j} c_i \, c_j \sum_{y,y'} g(x_i, y) \, g(x_j, y') \, k_{t-1}(y, y') \\
&= \sum_{y,y'} \left(\sum_i c_i \, g(x_i, y) \right) \times \\
&\qquad \left(\sum_j c_j \, g(x_j, y') \right) \, k_{t-1}(y, y') \\
&= \sum_{y,y'} \gamma_y \, \gamma_{y'} \, k_{t-1}(y, y') \\
&= \left\| \sum_y \gamma_y \Phi_{t-1}(y) \right\|_{\mathcal{H}} \geq 0. \quad \square
\end{aligned} \tag{6}
$$

Corollary 1 *K defined in (1) is also a Mercer kernel.*

Proof. the proof is straightforward for the particular case $n = m$. As $k_t(x_i^p, x_j^q) = \langle \Phi_t(x_i^p), \Phi_t(x_j^q) \rangle$, we can write $K(\mathcal{S}_p, \mathcal{S}_q) = \sum_{i,j} \langle \Phi_t(x_i^p), \Phi_t(x_j^q) \rangle = \langle \sum_i \Phi_t(x_i^p), \sum_j \Phi_t(x_j^q) \rangle$ and this corresponds to a dot product in some Hilbert space. The proof can be found in [27, 14] for the general case of finite subsets of any length. $\square$

2.3. Algorithm and Setting

The factor β, in k_t, acts as a scale parameter and it is selected using

$$
\beta \leftarrow \mathbb{E}_r \left[\mathbb{E}_{\{X_1^r, X_2^r : d(X_1^r, X_2^r) \leq \epsilon\}} \; [d(X_1^r, X_2^r)] \right] \tag{7}
$$

here $\mathbb{E}$ denotes the expectation and X_1^r (also X_2^r) denotes a random variable standing for samples in $\mathcal{S}_r$. The coefficient α controls the tradeoff between the alignment quality and the neighborhood criteria. It is selected by cross-validation and it should guarantee $k_t(x_i^p, x_j^q) \in [0, 1]$. If $A = \sup_{i,j} \sum_{k,\ell} g(x_i^p, x_k^p) \times g(x_j^q, x_\ell^q)$, α should then be selected in $[0, \frac{\beta}{2\,A}]$ (see. appendix).

Let $P_{i,j}$ denotes the i^{th} row of the j^{th} column of P. Consider P, Q as the intrinsic adjacency matrices of $\mathcal{S}_p$ and $\mathcal{S}_q$ respectively defined as $P_{i,k} = g(x_i^p, x_k^p)$, $Q_{j,\ell} = g(x_j^q, x_\ell^q)$. Let U denotes the unit matrix and consider $D_{i,j} = d(x_i^p, x_j^q)$, $\mu_{i,j}^{(t)} = k_t(x_i^p, x_j^q)$. Now, $\mu_{i,j}^{(t)}$ is iteratively found using Algorithm ("CDK") (see table 2) and converges to a fixed point (see. appendix).

3. Peformance

3.1. Databases

Experiments were conducted on the Swedish set (15 classes, 75 images per category) and a random subset of

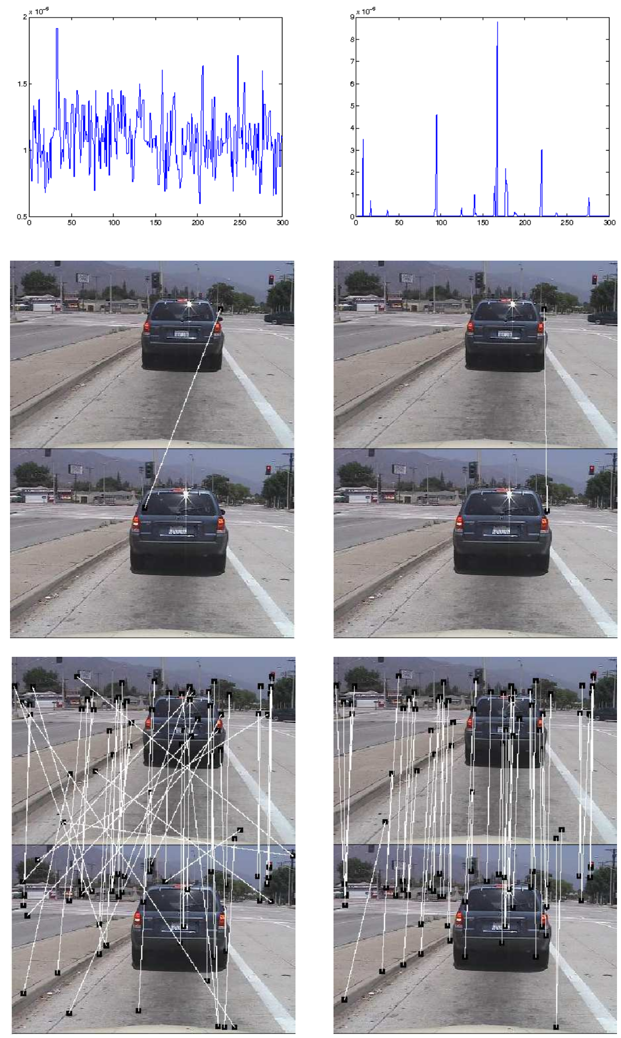

Figure 2. This figure shows a comparison of the matching results when using a naive matching strategy without geometry and our "context-dependent" kernel matching. (Top figures) show the distribution of the kernel values $k(x_i, x_j), j \in \mathcal{I}_q$ using a context-free kernel (left) and our "CDK" kernel (right). We can clearly see that the highest value changes its location so the matching results are now corrected (as shown in middle figures). (Bottom) other matching results.

Algorithm (CDK)

Initialization:

Set β using (7) and $\alpha \in [0, \frac{\beta}{2\,A}]$

Set $\mu^{(0)} \leftarrow k_0, t \leftarrow 0$

Repeat until $t \rightarrow T_{max}$ or $\|\mu^{(t)} - \mu^{(t-1)}\|_2 \rightarrow 0$

$$
\mu^{(t)} \leftarrow \exp\left(-D/\beta + \frac{2\alpha}{\beta} P \, \mu^{(t-1)} \, Q - U \right)
$$

Table 2. *The "CDK" kernel evaluation.*

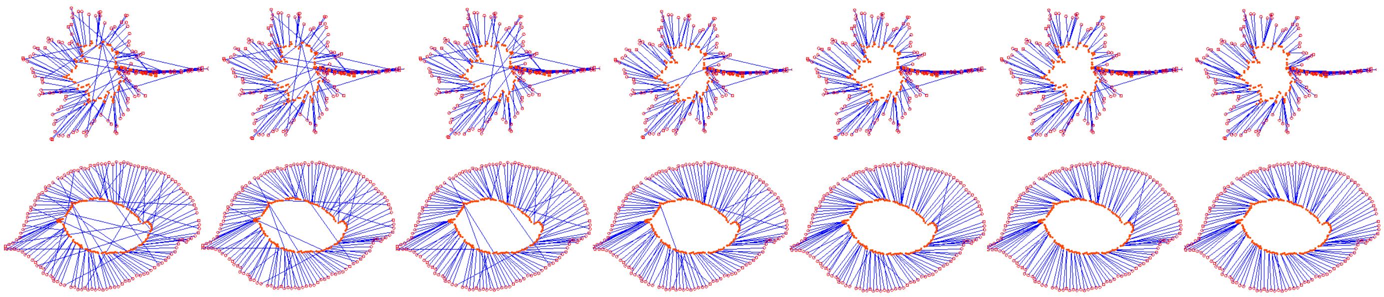

Figure 3. This figure shows the evolution of context-dependent silhouette matching on the Swedish set, for different and increasing values of α. We clearly see that when α increases the matching results are better. We set $\beta = 0.1$ and $t = 1$.

MNIST digit database (10 classes, 200 images per category). Each class in Swedish (resp. MNIST) is split into $50 + 25$ (resp. $100 + 100$) contours for training and testing. Interest points were sampled from each contour in MNIST (resp. Swedish) and encoded using the 60 (resp. 16) coefficients of the shape-context descriptor [4].

3.2. Generalization and Comparison

We evaluate $k_t,\ t \in \mathbb{N}^+$ using two initializations: (i) linear $k_0(x, x') = k_l(x, x') = \langle x, x' \rangle$ (ii) and polynomial $k_0(x, x') = k_p(x, x') = (\langle x, x' \rangle + 1)^2$. Our goal is to show the improvement brought when using $k_t,\ t \in \mathbb{N}^+$, so we tested it against the standard context-free kernels k_l and k_p (i.e., $k_t,\ t = 0$). For this purpose, we train a "one-versus-all" SVM classifier for each class in both MNIST and Swedish using the subset kernel $K(\mathcal{S}_p, \mathcal{S}_q) = \sum_{x \in \mathcal{S}_p, x' \in \mathcal{S}_q} k_t(x, x')$. The performance are measured, on different test sets, using n-fold cross-validation ($n = 5$).

We remind that β is set using (7) as the left-hand side of k_t corresponds to the Gaussian kernel with scale β. In practice, $\beta = 0.1$. The influence (and the performance) of the right-hand side of k_t increases as α increases (see. Figure 3), nevertheless and as shown in the appendix, the convergence of k_t to a fixed point is guaranteed only if $\alpha \in [0, \frac{\beta}{2\ A}]$. Therefore, it is obvious that α should be set to $\frac{\beta}{2\ A}$ where $A = \sup_{i,j} \sum_{k,\ell} g(x_i^p, x_k^p) \times g(x_j^q, x_\ell^q)$ (in practice, $0 \leq g \leq 1$ and $A = 1$).

Tables (3, 4), show the 5-fold cross validation errors on MNIST and Swedish for different iterations; we clearly see the out-performance and the improvement of the "CDK" kernel ($k_t, t \in \mathbb{N}^+$) with respect to the context-free kernels used for initialization ($k_0 = k_l$ and $k_0 = k_p$.)

INITIALIZATION	LINEAR	POLYNOMIAL
ITERATIONS (MNIST)		
k_0	11.4$\pm$ 4.42	9.15 $\pm$ 4.63
k_1	8.80$\pm$ 4.77	5.6 $\pm$ 2.72
k_2	6.90$\pm$ 3.55	5.8 $\pm$ 2.36
k_3	6.90$\pm$ 3.41	5.2 $\pm$ 2.07
k_4	**6.90$\pm$ 3.41**	**5.2 $\pm$ 2.07**

Table 3. The mean and the standard deviation of the 5-fold error on the MNIST database. Poly and Lin stand respectively for the polynomial and the linear kernels which are used as initialization of the "CDK" kernel. We can see a clear and a consistent gain through different iterations and also the convergence of the error.

INITIALIZATION	LINEAR	POLYNOMIAL
ITERATIONS (SWEDISH)		
k_0	11.7$\pm$ 2.88	6.53$\pm$ 6.34
k_1	6.00$\pm$ 2.30	3.33$\pm$ 2.73
k_2	**3.06$\pm$ 1.88**	**3.33$\pm$ 2.73**

Table 4. The same experiments are shown on the Swedish set.

4. Remarks and Discussion

The adjacency matrix P, in k_t, provides the intrinsic properties and also characterizes the geometry of an object $\mathcal{S}_p$. Let us remind $\mathcal{N}_p(x_i^p) = \{x_k^p \in \mathcal{S}_p : k \neq i,\ \|\psi_g(x_i^p) - \psi_g(x_k^p)\|_2 \leq \epsilon_p\}$ and $P_{i,j} = \mathbb{1}_{\{x_j^q \in \mathcal{N}_p(x_i^p)\}}$. It is easy to see that P is translation and rotation invariant and can also be made scale invariant when ϵ_p is adapted to the scale of $\psi_g(x_i^p)$. 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 k_t involves similarity invariant descriptors $\psi_f(x_i^p)$, $\psi_f(x_j^q)$ so k_t (and K) is similarity invariant.

The out-performance of our kernel comes essentially from the inclusion of the context. This strongly improves the precision and helps including the intrinsic properties (geometry) of objects. Even though tested only on visual object recognition, our kernel can be extended to many other pattern analysis problems such as bioinformatics, speech and text. For instance, in text analysis and particularity machine translation [28], 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) [16].

Finally, one current limitation of our kernel k_t resides in its evaluation complexity. Assuming k_{t-1} known, for a given pair x_i^p, x_j^q, this complexity is $O\left(\max(N^2, s)\right)$, where s is the dimension of $\psi_f(x_i^p)$ and $N = \max_{i,p} \#\{\mathcal{N}_p(x_i^p)\}$. 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 $k_t(x_i^p, x_j^q)$ 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 [31]) will reduce this complexity to $O(s)$.

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. The latter is achieved for shape recognition problems and has better performance than SVM with context-free kernels. Future work includes the comparison of our kernel with other context-free kernels and its application in scene and object understanding using more challenges and databases.

Appendix

Proposition 1 (cont.)

Proof. Let us consider $\mu = \left\{k(x_i^p, x_j^q) = \exp(-U_{ij}^2), U_{ij} \in \mathbb{R},\ i \in \mathcal{I}_p, j \in \mathcal{I}_q\right\}$, and $U = \{U_{ij}\}$. Per definition the bounds on $\{k(x_i^p, x_j^q)\}$ are satisfied. Now, the objective function (3) can be rewritten as

$$\begin{aligned} \min_U \quad & \sum_{i,j} \exp(-U_{ij}^2)\, d(x_i^p, x_j^q) \;-\; \beta \sum_{i,j} \exp(-U_{ij}^2)\, U_{ij}^2 \;- \\ & \alpha \sum_{i,j} \sum_{k,\ell} \mathbb{V}(x_i^p, x_k^p, x_j^q, x_\ell^q)\ \exp(-U_{ij}^2)\ \exp(-U_{kl}^2) \\ \text{s.t.} \quad & \sum_j \exp(-U_{ij}^2) = 1, \quad \forall\, i \in \mathcal{I}_p \end{aligned} \tag{8}$$

By introducing Lagrange coefficients λ for the equality constraint $\{\sum_{i,j} \exp(-U_{ij}^2) = 1\}$, the above constrained minimization problem can now be rewritten:

$$\begin{aligned} \min_{U,\lambda} \quad & \mathcal{L}(U,\lambda) = \\ \min_{U,\lambda} \quad & \sum_{i,j} \exp(-U_{ij}^2)\, d(x_i^p, x_j^q) \;-\; \beta \sum_{i,j} \exp(-U_{ij}^2)\, U_{ij}^2 - \\ \alpha \quad & \sum_{i,j,k,\ell} \exp(-U_{ij}^2)\ \exp(-U_{kl}^2)\ \mathbb{V}(x_i^p, x_k^p, x_j^q, x_\ell^q) + \\ \lambda \quad & \left(\sum_{i,j} \exp(-U_{ij}^2) - 1\right) \end{aligned} \tag{9}$$

The conditions for optimality, i.e., when the gradient with respect to $\{U_{ij}\}$ and λ vanishes, lead to :

$$\begin{aligned} & -2\, U_{ij}\ \exp(-U_{ij}^2)\, d(x_i^p, x_k^p) \;+\; 2\,\beta\, U_{ij}^3 \exp(-U_{ij}^2) \;- \\ & 2\,\beta\, U_{ij}\ \exp(-U_{ij}^2) \;-\; 2\lambda\, U_{ij}\ \exp(-U_{ij}^2) + \\ & 4\,\alpha\ \mathbb{V}(x_i^p, x_i^p, x_j^q, x_j^q)\, U_{ij}\ \exp(-U_{ij}^2) + \\ & 4\,\alpha \sum_{k,\ell} \mathbb{V}(x_i^p, x_k^p, x_j^q, x_\ell^q)\, U_{ij} \exp(-U_{ij}^2) \exp(-U_{kl}^2) \;=\; 0 \end{aligned} \tag{10}$$

and

$$\sum_{i,j} \exp(-U_{ij}^2) \;=\; 1$$

$\dfrac{\partial \mathcal{L}}{\partial U_{ij}} = 0$ implies:

$$\begin{aligned} & -d(x_i^p, x_j^q) \;+\; \beta\ \left(U_{ij}^2 - 1\right) \;-\; \lambda \;+\; 2\alpha\, \mathbb{V}(x_i^p, x_i^p, x_j^q, x_j^q) \;- \\ & 2\alpha \left(\sum_{k,\ell} \mathbb{V}(x_i^p, x_k^p, x_j^q, x_\ell^q) e^{-U_{kl}^2}\right) = 0 \end{aligned} \tag{11}$$

so $k(x_i^p, x_j^q)$ is equal to

$$\begin{aligned} & \exp\left(-U_{ij}^2\right) \\ = \quad & \exp\left(-\frac{d(x_i^p, x_j^q)}{\beta}\right)\ \exp(-1)\ \times \\ & \exp\left(\frac{2\alpha}{\beta} \sum_{k,\ell} \mathbb{V}(x_i^p, x_k^p, x_j^q, x_\ell^q)\, k(x_k^p, x_\ell^q)\right)\ \times \\ & \exp\left(-\frac{2\alpha}{\beta} \mathbb{V}(x_i^p, x_i^p, x_j^q, x_j^q)\right)\ \exp\left(-\frac{\lambda}{\beta}\right) \end{aligned} \tag{12}$$

It is easy to see that $\exp\left(-\dfrac{2\alpha}{\beta} \mathbb{V}(x_i^p, x_i^p, x_j^q, x_j^q)\right)$ is constant (i.e., independent from i, j). Now $\dfrac{\partial \mathcal{L}}{\partial \lambda} = 0$, implies $\exp\left(-\frac{\lambda}{\beta}\right) = \exp(1) / \sum_{i,j} Z_{ij}$

with $Z_{ij} = \exp\left(-\dfrac{d(x_i^p, x_j^q)}{\beta} \;+\; \dfrac{2\alpha}{\beta} \sum_{k,\ell} \mathbb{V}(x_i^p, x_k^p, x_j^q, x_\ell^q)\, k(x_k^p, x_\ell^q)\right)$

By plugging the above two equations into (12), the global form of the solution $\{k_t(x_i^p, x_j^q)\}$ which minimizes the constrained minimization problem (3) is:

$$\frac{1}{Z_t} \times \exp\left(-\frac{d(x_i^p, x_j^q)}{\beta}\right) \times \exp\left(\frac{2\alpha}{\beta} \sum_{k,\ell} \mathbb{V}(x_i^p, x_k^p, x_j^q, x_\ell^q) k_{t-1}(x_k^p, x_\ell^q)\right) \tag{13}$$

where $Z_t = \sum_{i,j} Z_{ij}^{(t)}$. The solution of (3) corresponds to a fixed-point which is found iteratively $\square$

Convergence

Let us assume $0 \le g \le 1$, and remind $\mu^{(t)} \in \mathbb{R}^{n \times m}$ be the vector of components $\mu_{i,j}^{(t)} = k_t(x_i^p, x_j^q)$. Introduce the mapping $f : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^{n \times m}$ defined by its component

$$f_{i,j}(v) = \exp\left(-1 - \frac{d(x_i^p, x_j^q)}{\beta} + \frac{2\alpha}{\beta} \sum_{k,\ell} g(x_i^p, x_k^p) g(x_j^q, x_\ell^q) v_{k,\ell}\right) \tag{14}$$

By construction of the kernel k_t, we have $\mu^{(t)} = f\big(\mu^{(t-1)}\big)$. Let A and B satisfy

$$\sup_{1 \le i \le n,\ 1 \le j \le m} \sum_{k,\ell} g(x_i^p, x_k^p) g(x_j^q, x_\ell^q) \le A \tag{15}$$

$$\sum_{i,j} \exp\left(-1 - \frac{d(x_i^p, x_j^q)}{\beta}\right) \le B \tag{16}$$

Consider $L = \frac{2B\alpha}{\beta} \exp\big(\frac{2\alpha A}{\beta}\big)$, and let $\mathcal{B} = \big\{v \in \mathbb{R}^{n \times m} : \forall\, 1 \le i \le n,\ 1 \le j \le m,\ |v_{i,j}| \le 1\big\}$ be the $\|\cdot\|_\infty$-ball of radius 1. Finally, let $\|\cdot\|_1$ denote the 1-norm on $\mathbb{R}^{n \times m}$: $\|u\|_1 = \sum_{1 \le i \le n,\ 1 \le j \le n} |u_{i,j}|$.

Proposition 4 *If $\|\mu^{(0)}\|_\infty \le 1$ and $2\alpha A \le \beta$, then we have $f(\mathcal{B}) \subset \mathcal{B}$, and on $\mathcal{B}$, f is L-Lipschitz for the norm $\|\cdot\|_1$.*

In particular, if $L < 1$, then there exists a unique $\tilde{v} \in \mathcal{B}$ such that $f(\tilde{v}) = \tilde{v}$, and the sequence $(\mu^{(t)})$ satisfies

$$\|\mu^{(t)} - \tilde{v}\|_1 \le L^t \|\mu^{(0)} - \tilde{v}\|_1 \underset{t \to +\infty}{\longrightarrow} 0. \tag{17}$$

Proof. The first assertion is proved by induction by checking that for $\|v\|_\infty \le 1$, we have

$$f_{i,j}(v) \le \exp\left(-1 + \frac{2\alpha}{\beta} \sum_{k,\ell} g(x_i^p, x_k^p) g(x_j^q, x_\ell^q) v_{k,\ell}\right) \tag{18}$$

$$\le \exp\left(-1 + \frac{2\alpha}{\beta} A\right) \le 1. \tag{19}$$

For the second assertion, note that for any v in $\mathcal{B}$, we have $\big|\frac{\partial f_{i,j}}{\partial v_{k,\ell}}(v)\big| \le \exp\big(-1 - \frac{d(x_i^p, x_j^q)}{\beta}\big)$ For any v, v' in $\mathcal{B}$, we have

$$\|f(v) - f(v')\|_1 = \sum_{i,j} |f_{i,j}(v) - f_{i,j}(v')| = (**)$$

$$(**) \le \sum_{i,j} \exp\big(-1 - \frac{d(x_i^p, x_j^q)}{\beta}\big) \frac{2\alpha}{\beta} \exp\big(\frac{2\alpha}{\beta} A\big) \tag{20}$$

$$\times \big| \sum_{k,\ell} g(x_i^p, x_k^p) g(x_j^q, x_\ell^q) v_{k,\ell} \tag{21}$$

$$- \sum_{k,\ell} g(x_i^p, x_k^p) g(x_j^q, x_\ell^q) v'_{k,\ell} \big| \tag{22}$$

$$\le \sum_{i,j} \exp\big(-1 - \frac{d(x_i^p, x_j^q)}{\beta}\big) \frac{2\alpha}{\beta} \exp\big(\frac{2\alpha}{\beta} A\big) \|v - v'\|_1 \tag{23}$$

$$\le L\|v - v'\|_1 \tag{24}$$

which proves the second assertion. The last assertion directly comes from the fixed-point theorem □.

References

[1] 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.

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

[3] 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.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. *In NIPS*, 2000.

[5] C. Bishop. Pattern recognition and machine learning. *Springer*, 2007.

[6] 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.

[7] S. Boughorbel, J. Tarel, and N. Boujemaa. The intermediate matching kernel for image local features. *IEEE International Joint Conference on Neural Networks*, 2005.

[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] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. *http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html*, 2007.

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

[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 UCS-CRL-99-10, UC Santa Cruz*, 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] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, 2003.

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

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

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

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

[21] 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.

[22] E. 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.

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

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

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

[27] J. Shawe-Taylor and N. Cristianini. Support vector machines and other kernel-based learning methods. *Cambridge University Press*, 2000.

[28] 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.

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

[30] V. N. Vapnik. Statistical learning theory. *A Wiley-Interscience Publication*, 1998.

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

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

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