Explicit Context-Aware Kernel Map Learning for Image Annotation

Hichem SAHBI

CNRS - TELECOM ParisTech
hichem.sahbi@telecom-paristech.fr

Abstract. *In kernel methods, such as support vector machines, many existing kernels consider similarity between data by taking into account only their content and without context. In this paper, we propose an alternative that upgrades and further enhances usual kernels by making them context-aware. The proposed method is based on the optimization of an objective function mixing content, regularization and also context. We will show that the underlying kernel solution converges to a positive semi-definite similarity, which can also be expressed as a dot product involving "explicit" kernel maps. When combining these context-aware kernels with support vector machines, performances substantially improve for the challenging task of image annotation.*

Keywords: Context-Aware Kernels, Explicit Kernel Maps, SVMs.

1 Introduction

Image annotation is a major challenge in computer vision, which consists in assigning a list of keywords to a given image [14, 20, 6]. These keywords may either correspond to physical entities (pedestrians, cars, etc.) or to high level concepts resulting from the interaction of many entities into scenes (races, fights, etc.). In both cases, image annotation is challenging due to the perplexity when assigning keywords to images especially when the number of possible keywords is taken from a large vocabulary and when analyzing highly semantic contents.

Existing annotation methods usually model image observations using low level features (color, texture, shape, etc.), and then assign keywords to these observations using a variety of machine learning and inference techniques such as latent Dirichlet allocation [2], hidden Markov models [14], probabilistic latent semantic analysis [17] and support vector machines (SVMs) [7]. These learning machines are used to model correspondences between keywords and low level features and make it possible to assign keywords to new images. Among learning techniques those based on kernel methods, mainly SVMs, are particularly successful [1] but their success remains highly dependent on the choice of kernels. The latter, defined as symmetric and positive semi-definite functions [23], should reserve large values to very similar contents and vice-versa.

[1] See for instance the periodic and the challenging ImageCLEF benchmark [20].

Considering a collection of images, each one seen as a constellation of primitives (eg. interest points) [8]. Two families of kernels were introduced, in the literature, in order to handle these types of data; *holistic* and *alignment-based* kernels. Holistic kernels first map constellations of primitives to feature vectors, by estimating their first or high order statistics or by aggregating them [13, 18]. Then, similarity is defined as any decreasing function of a distance between these feature vectors, via usual kernels (such as gaussian or histogram intersection). Note that the resulting kernels are positive semi-definite per construction. In the second family of kernels, methods proceed differently [1, 9, 5, 16, 27] and consider similarity proportional to the quality of aligning primitives. In contrast to the first family, the positive definiteness of this second family of kernels is not straightforward and not always guaranteed. Notice also that holistic kernels are naturally more flexible and invariant to geometric transformations, but alignment-based kernels are more discriminating; it is clear that kernels that gather the advantages of the two aforementioned families of kernels are preferred.

We are interested, in this work, in the integration of context in kernels in order to further enhance their discrimination power while keeping their flexibility to handle constellations of primitives, their invariance to geometric transformations and also their efficiency. Context is important and has, indeed, played an important role in leveraging the performances of many computer vision tasks, and mainly those based on Markov models (see for instance [11, 25, 12, 22, 19]), but the novel part of this work aims to integrate context, in kernel design useful for classification and annotation, and plug these kernels in support vector machines in order to take benefit from their well established generalization power [26]. Again, given a collection of images, each one described as a constellation of interest points, the proposed method is based on the optimization of an objective function mixing a fidelity term, a context criterion and a regularization term. The fidelity term, takes into account the visual content of interest points in order to measure the quality of their alignments, so high quality alignments encourage high kernel values. The context criterion, considers the global scene structure and allows us to further enhance the relevance of our designed kernel, by restoring the similarity *iff* pairs of aligned interest points are *also* surrounded by good quality alignments that should also share the same context (see § 2.1). The regularization term controls the smoothness of the learned kernel and makes it possible to obtain a closed form solution.

Note that this work is built upon [24] but includes many differences (see § 2):
- A simplification of the learning model; which now includes an *unconstrained* objective function, easier-to-solve. Furthermore, the number of parameters of our model reduces now to one, and corresponds to the weight of context.
- A new study of the theoretical properties of our solution; mainly (i) the introduction of a loose upper-bound, about the weight of context, that guarantees convergence of the learned kernel to a fixed-point and (ii) the study of the positive definiteness which shows that the obtained kernel solution can be written as a dot product, involving "explicit" kernel maps. Indeed, in spite of the non-linearity of our kernel, its map is explicit, so one may use extremely fast SVM

solvers in order to handle large scale datasets[2] and without the overhead of pre-computing gram matrices and solving quadratic programming (QP) problems.

2 Explicit Context-Aware Kernel Map Learning

Let $\{\mathcal{I}_p\}_p$ be a collection of images and let $\mathcal{S}_p = \{\mathbf{x}_1^p, \ldots, \mathbf{x}_n^p\}$ be the list of interest points of an image $\mathcal{I}_p$ (the value of n may vary with the image $\mathcal{I}_p$). The set $\mathcal{X}$ of all possible interest points is the union over all possible images of $\{\mathcal{I}_p\}$: $\mathcal{X} = \cup_p \mathcal{S}_p$. Consider $\mathcal{S}_p, \mathcal{S}_q \subseteq \mathcal{X}$ as two finite subsets of $\mathcal{X}$, the convolution kernel $\mathcal{K}$, between $\mathcal{S}_p = \{\mathbf{x}_i^p\}_{i=1}^n$ and $\mathcal{S}_q = \{\mathbf{x}_j^q\}_{j=1}^{n'}$, is defined as $\mathcal{K}(\mathcal{S}_p, \mathcal{S}_q) = \sum_{i,j} \kappa\left(\mathbf{x}_i^p, \mathbf{x}_j^q\right)$, here κ may be any symmetric and continuous function on $\mathcal{X} \times \mathcal{X}$, so $\mathcal{K}$ will also be continuous and symmetric, and if κ is positive semi-definite (p.s.d) then $\mathcal{K}$ will also be p.s.d [10]. Since $\mathcal{K}$ is defined as the sum of all the pairwise 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 $\kappa\left(\mathbf{x}_i^p, \mathbf{x}_j^q\right)$ should ideally be high only if $\mathbf{x}_i^p$ actually matches $\mathbf{x}_j^q$, so κ needs to be appropriately designed while being p.s.d.
Formally, an interest point $\mathbf{x}$ is defined as $\mathbf{x} = (\psi_g(\mathbf{x}), \psi_o(\mathbf{x}), \psi_s(\mathbf{x}), \psi_f(\mathbf{x}), \omega(\mathbf{x}))$ where the symbol $\psi_g(\mathbf{x}) \in \mathbb{R}^2$ stands for the 2D coordinates of $\mathbf{x}$ while the orientation and scale of $\mathbf{x}$ (respectively denoted $\psi_o(\mathbf{x}) \in [-\pi, +\pi]$ and $\psi_s(\mathbf{x}) \in]0, \max]$) are provided by the SIFT gradient and scale respectively. We have an extra information about the visual content or features of $\mathbf{x}$ (denoted $\psi_f(\mathbf{x}) \in \mathbb{R}^s$); in our case, these visual features result from the concatenation of (i) 128 SIFT coefficients; [15], (ii) 3-channel color histograms, of 20 dimensions each and (iii) shape context dartboard [3] of 8 bands and 8 sectors; both (ii) and (iii) are computed locally, in a disk centered at $\mathbf{x}$ with a radius proportional to $\psi_s(\mathbf{x})$. We also use $\omega(\mathbf{x})$ to denote the image from which the interest point comes from, so that two interest points with the same location, feature, scale 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).
Introduce the context of $\mathbf{x}$, $\mathcal{N}^{\theta,\rho}(\mathbf{x}) = \{\mathbf{x}' : \omega(\mathbf{x}') = \omega(\mathbf{x}), \mathbf{x}' \neq \mathbf{x} \text{ s.t. (1) holds}\}$,

$$\|\psi_g(\mathbf{x}) - \psi_g(\mathbf{x}')\|_2 \in \left[\frac{\rho-1}{N_r}\,\epsilon_p, \frac{\rho}{N_r}\,\epsilon_p\right], \text{angle}\big(\psi_o(\mathbf{x}), \psi_g(\mathbf{x}') - \psi_g(\mathbf{x})\big) \in \left[\frac{\theta-1}{N_a}\pi, \frac{\theta}{N_a}\pi\right]. \tag{1}$$

Here ϵ_p is the radius of a neighborhood disk surrounding $\mathbf{x}$ and $\theta = 1, ..., N_a$, $\rho = 1, ..., N_r$ correspond to indices of different parts of that disk. In practice, N_a and N_r correspond to 8 sectors and 8 bands. In the remainder of this paper, $\mathcal{N}^{\theta,\rho}(\mathbf{x})$ will simply be denoted as $\mathcal{N}^c(\mathbf{x})$, with $c = (\theta - 1)N_r + \rho$.

2.1 The Method

We can view a kernel κ on $\mathcal{X}$ as a matrix $\mathbf{K}$ in which the "$(\mathbf{x}, \mathbf{x}')-$element" is the similarity between $\mathbf{x}$ and $\mathbf{x}'$: $\mathbf{K}_{\mathbf{x},\mathbf{x}'} = k(\mathbf{x}, \mathbf{x}')$. Let $\mathbf{P}_c$ be the intrinsic adjacency

[2] Such as stochastic gradient descent [4]. When the kernel map is explicit, the complexity of this method is linear in the size of the training data instead of quadratic.

matrix defined as $\mathbf{P}_{c,\mathbf{x},\mathbf{x}'} = g_c(\mathbf{x},\mathbf{x}')$, where g is a decreasing function of any (pseudo) distance involving $(\mathbf{x},\mathbf{x}')$, not necessarily symmetric. In practice, we consider $g_c(\mathbf{x},\mathbf{x}') = \frac{1}{m}\mathbb{1}_{\{\mathbf{x}'\in\mathcal{N}^c(\mathbf{x})\}}$, with $m = |\mathcal{X}|$. Let $\mathbf{S}$ be a matrix with $\mathbf{S}_{\mathbf{x},\mathbf{x}'} = \langle\psi_f(\mathbf{x}),\psi_f(\mathbf{x}')\rangle$. We propose to use the kernel on $\mathcal{X}$ defined by solving

$$\min_{\mathbf{K}} tr(-\mathbf{K}\mathbf{S}')) \; - \alpha \sum_c tr(\mathbf{K}\mathbf{P}_c\mathbf{K}'\mathbf{P}_c') + \frac{\beta}{2}\|\mathbf{K}\|_2^2, \tag{2}$$

with $\alpha,\beta \geq 0$, $'$, tr denote matrix transpose and trace operator respectively. The first term, in the above optimization problem, measures the quality of matching two features $\psi_f(\mathbf{x})$, $\psi_f(\mathbf{x}')$ (this is considered as the inner product, $\langle\psi_f(\mathbf{x}),\psi_f(\mathbf{x}')\rangle$, between the visual features of $\mathbf{x}$ and $\mathbf{x}'$). A small value of this inner product should result into a small value of $\kappa(\mathbf{x},\mathbf{x}')$ and vice-versa. The second term is a neighborhood (or context) criterion which considers that a high value of $\kappa(\mathbf{x},\mathbf{x}')$ should imply high kernel values in the neighborhoods $\mathcal{N}^c(\mathbf{x})$ and $\mathcal{N}^c(\mathbf{x}')$. This criterion also makes it possible to consider the spatial configuration of the neighborhood of each interest point in the matching process. The third term is a regularization criterion that controls the smoothness of the learned kernel and also helps getting a closed form kernel solution.

Proposition 1. *Let $\gamma = \alpha/\beta$ and $\|.\|_1$ denote the entrywise L_1-norm. Provided that the following inequality holds,*

$$\gamma < \big\| \sum_c \mathbf{P}_c \; \mathbf{1}_{mm} \; \mathbf{P}'_c \big\|_1^{-1} \tag{3}$$

the optimization problem (2) admits a unique solution $\tilde{\mathbf{K}}$ as the limit of

$$\mathbf{K}^{(t+1)} = \psi\big(\mathbf{K}^{(t)}\big), \tag{4}$$

here $\psi : \mathbb{R}^{m\times m} \to \mathbb{R}^{m\times m}$ is defined as $\psi(\mathbf{K}) = \mathbf{S} + \gamma\sum_c \mathbf{P}_c \; \mathbf{K} \; \mathbf{P}'_c$, and $\mathbf{1}_{mm}$ is a $m\times m$ square matrix of ones. Furthermore, the kernels $\mathbf{K}^{(t)}$ in (4) satisfy the convergence property: $\big\|\mathbf{K}^{(t)} - \tilde{\mathbf{K}}\big\|_1 \leq L^t \big\|\mathbf{K}^{(0)} - \tilde{\mathbf{K}}\big\|_1$, with $L = \gamma\|\sum_c \mathbf{P}_c \; \mathbf{1}_{mm} \; \mathbf{P}'_c\|_1$ and $\mathbf{K}^{(0)} = \mathbf{S}$.

Proof. See appendix.

Now, we will show how explicit kernel maps can be obtained from this solution.

2.2 Explicit p.s.d Kernels

Definition 1. *Let κ be symmetric and continuous similarity function. κ is referred to as explicit p.s.d kernel if κ is p.s.d (i.e. $\exists\phi : \; \forall\mathbf{x},\mathbf{x}' \in \mathcal{X}$, $\kappa(\mathbf{x},\mathbf{x}') = \langle\phi(\mathbf{x}),\phi(\mathbf{x}')\rangle$) and its mapping ϕ is explicit and finite dimensional.*

Example. following the above definition, the polynomial kernel defined, between $\mathbf{x}_a = (a_1 \;\; a_2)^t$, $\mathbf{x}_b = (b_1 \;\; b_2)^t$, as $\kappa(\mathbf{x}_a,\mathbf{x}_b) = \langle\mathbf{x}_a,\mathbf{x}_b\rangle^2$, is explicit p.s.d since $\langle\mathbf{x}_a,\mathbf{x}_b\rangle^2 = \langle\phi(\mathbf{x}_a),\phi(\mathbf{x}_b)\rangle$, with $\phi(\mathbf{x}_a) = (a_1^2 \;\; \sqrt{2}a_1a_2 \;\; a_2^2)^t$ and $\phi(\mathbf{x}_b) = (b_1^2 \;\; \sqrt{2}b_1b_2 \;\; b_2^2)^t$, while the gaussian kernel $\kappa(\mathbf{x}_a,\mathbf{x}_b) = \exp(-\frac{1}{\sigma}\|\mathbf{x}_a - \mathbf{x}_b\|_2^2)$ is p.s.d but not explicit p.s.d as its mapping is infinite dimensional.

Proposition 2. *The similarity functions* $\mathbf{K}^{(t+1)}_{\mathbf{x},\mathbf{x}'}$, $(t = 0, 1, \dots)$ *defined, in proposition (1), as* $\mathbf{K}^{(t+1)}_{\mathbf{x},\mathbf{x}'} = \left(\mathbf{S} + \gamma \sum_c \mathbf{P}_c \ \mathbf{K}^{(t)} \ \mathbf{P}'_c\right)_{\mathbf{x},\mathbf{x}'}$ *are explicit p.s.d kernels.*

Proof. See appendix.

Algorithm (1) shows the iterative process of kernel map learning. According to this algorithm and the previous proposition, it is clear that the mapping $\mathbf{\Phi}^{(t+1)}$ is not equal to $\mathbf{\Phi}^{(t)}$ since the dimensionality of the map increases w.r.t. t. However, the convergence of the inner product $\mathbf{\Phi}'^{(t+1)}\mathbf{\Phi}^{(t+1)}$ to a fixed point is guaranteed when (3) is satisfied, i.e., the gram matrices of the designed kernel maps are convergent.
Resulting from the definition of the adjacency matrices $\{\mathbf{P}_c\}$, in (2.1), it is easy to see that the latter are block diagonal so learning kernel maps could be achieved image per image with obviously the same number of iterations, i.e., the evaluation of kernel maps of a given image is independent from others and hence not transductive; and this makes it incremental. Now, considering $\tilde{\mathbf{K}}$ as the limit of $\psi(\mathbf{K})$, the new form of the convolution kernel $\mathcal{K}$ between two sets of interest points $\mathcal{S}_p$, $\mathcal{S}_q$ can be rewritten $\mathcal{K}(\mathcal{S}_p, \mathcal{S}_q) = \sum_{(\mathbf{x},\mathbf{x}')\in\mathcal{S}_p\times\mathcal{S}_q} \langle \tilde{\mathbf{\Phi}}_{\mathbf{x}}, \tilde{\mathbf{\Phi}}_{\mathbf{x}'} \rangle$, again $\tilde{\mathbf{\Phi}}$ is the learned kernel map obtained after convergence of algorithm (1) and the subscript in $\tilde{\mathbf{\Phi}}_{\mathbf{x}}$ denotes the restriction of this map to an interest point $\mathbf{x}$. It is easy to see that $\mathcal{K}$ is an explicit p.s.d kernel as it can be rewritten as a dot product involving finite dimensional and explicit maps, i.e., $\mathcal{K}(\mathcal{S}_p, \mathcal{S}_q) = \langle \phi_{\mathcal{K}}(\mathcal{S}_p), \phi_{\mathcal{K}}(\mathcal{S}_q) \rangle$, with $\phi_{\mathcal{K}}(\mathcal{S}_p) = \sum_{\mathbf{x}\in\mathcal{S}_p} \tilde{\mathbf{\Phi}}_{\mathbf{x}}$, which clearly shows that each constellation of interest points $\mathcal{S}_p$ can be indexed simply with the explicit kernel map $\phi_{\mathcal{K}}(\mathcal{S}_p)$.

Algorithm 1: Recursive kernel map learning

Input: The union of all interest points $\{\mathbf{x}_i\}$ in $\mathcal{X}$.
Output: Learned kernel maps $\tilde{\mathbf{\Phi}}$.

Set $t = 0$, γ using condition (3) and set the adjacency matrices $\{\mathbf{P}_c\}$ and $\mathbf{\Phi}^{(0)}$ with $\mathbf{\Phi}^{(0)}_{\mathbf{x}_i} = \psi_f(\mathbf{x}_i)$

repeat

$\mathbf{\Phi}^{(t+1)} \longleftarrow \left(\mathbf{\Phi}'^{(0)} \quad \gamma^{\frac{1}{2}} \ \mathbf{P}_1\mathbf{\Phi}'^{(t)} \ \dots \quad \gamma^{\frac{1}{2}}\mathbf{P}_{N_rN_a}\mathbf{\Phi}'^{(t)} \right)'$,

Set $t \leftarrow t + 1$

until $\left\|\mathbf{\Phi}'^{(t+1)}\mathbf{\Phi}^{(t+1)} - \mathbf{\Phi}'^{(t)}\mathbf{\Phi}^{(t)}\right\|_1 \rightsquigarrow 0$ *or* $t > T_{max}$;

Set $\tilde{\mathbf{\Phi}} \leftarrow \mathbf{\Phi}^{(t)}$

3 Experiments

We plugged the learned kernel $\mathcal{K}$ into SVMs in order to evaluate its performance. The targeted task is image annotation; given a picture of a database, the goal is to predict which concepts (classes) are present into that picture. For this purpose, we trained "one-versus-all" SVM classifiers for each concept; we use three random folds (75% of a database) for SVM training and the remaining fold for

Fig. 1. Sample of images from Swedish leaf (left) and ImageCLEF (right) databases.

testing. We repeat this training process through different folds, for each concept, and we take the average equal error rates (EERs) of the underlying SVM classifiers. This makes classification results less sensitive to sampling.
We run these experiments on different databases ranging from simple ones such as the Olivetti face database to relatively more challenging ones such as the Swedish and the extremely challenging ImageCLEF Photo Annotation database (see Fig. 1). The latter contains $18,000$ pictures split into 93 categories; a subset of $8,000$ images was used for training and testing as ground truth was publicly available for this subset only. The Swedish database contains 15 leaf species, each one represented by 75 examples, resulting into $1,125$ images while the Olivetti set is a well known face database of 40 persons each one contains 10 instances. For each image in these databases, we run the SIFT detector [15] in order to extract a constellation of interest points, and each one is described with the visual features discussed in Section (2).

Our goal is to show the improvement brought when using the learned kernel maps $\{\mathbf{\Phi}^{(t)}\}_{t\in\mathbb{N}^+}$, so we tested them against context-free kernel maps (i.e., $\mathbf{\Phi}^{(t)},\ t=0$). For that purpose, we trained the "one-versus-all" SVM classifiers for each class in Olivetti, Swedish and ImageCLEF sets using the convolution kernel $\mathcal{K}(\mathcal{S}_p,\mathcal{S}_q)=\big\langle\sum_{\mathbf{x}\in\mathcal{S}_p}\mathbf{\Phi}^{(t)}_{\mathbf{x}},\sum_{\mathbf{x}'\in\mathcal{S}_q}\mathbf{\Phi}^{(t)}_{\mathbf{x}'}\big\rangle$. The influence (and the performance) of the context term in $\mathbf{\Phi}^{(t)}$ (and hence $\mathbf{K}^{(t)}$) increases as γ increases (see example in Fig. 2), nevertheless and as shown earlier, the convergence of $\mathbf{K}^{(t)}$ to a fixed point is guaranteed only if Eq. (3) is satisfied. Intuitively, the parameter γ should then be relatively high while also satisfying condition (3). Larger values of γ (in practice $\gamma>1$), do not always guarantee convergence of the learned kernels and the classification performances may not converge to the best ones.
Table. (1) shows EERs of different baseline kernel maps and their upgraded context-aware versions. These baselines include: Linear: $\mathbf{K}^{(0)}_{\mathbf{x},\mathbf{x}'}=\langle\psi_f(\mathbf{x}),\psi_f(\mathbf{x}')\rangle$, Polynomial: $\langle\psi_f(\mathbf{x}),\psi_f(\mathbf{x}')\rangle^2$, RBF: $\exp\big(-\|\psi_f(\mathbf{x})-\psi_f(\mathbf{x}')\|_2^2/0.1\big)$, Triangular: $-\|\psi_f(\mathbf{x})-\psi_f(\mathbf{x}')\|_2$, Histogram intersection: $\sum_i\min(\psi_f(\mathbf{x})_i,\psi_f(\mathbf{x}')_i)$ and Chi-square: $1-\frac{1}{2}\sum_i\frac{(\psi_f(\mathbf{x})_i-\psi_f(\mathbf{x}')_i)^2}{(\psi_f(\mathbf{x})_i+\psi_f(\mathbf{x}')_i)}$. Each initialization $\mathbf{K}^{(0)}_{\mathbf{x},\mathbf{x}'}$ should be expressed as an explicit dot product $\langle\mathbf{\Phi}^{(0)}_{\mathbf{x}},\mathbf{\Phi}^{(0)}_{\mathbf{x}'}\rangle$. As this explicit kernel map is not available for the above kernels (excepting the linear), we apply kernel principal component analysis (KPCA) in order to decompose the matrix $\mathbf{K}^{(0)}$ into $\mathbf{\Phi}'^{(0)}\mathbf{\Phi}^{(0)}$ and hence approach these maps, i.e., each $\mathbf{x}$ is mapped to a finite dimensional vector

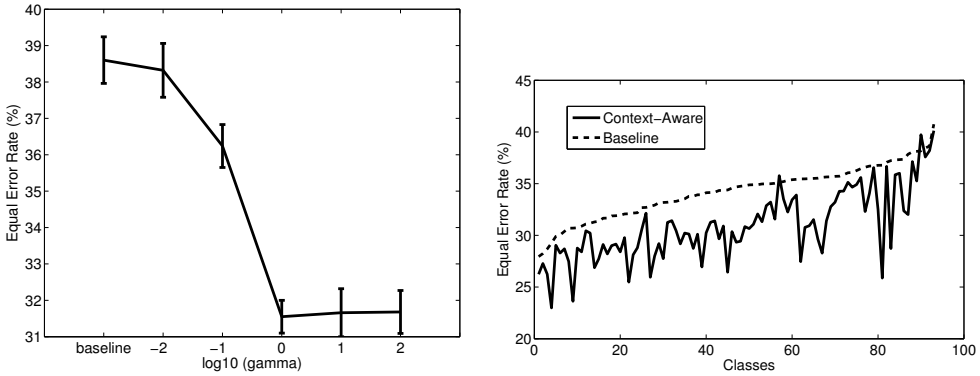

Fig. 2. The left-hand side figure shows the evolution of the overall (average) EER + standard deviation, of our context-aware kernel (with $N_a = N_r = 8$ and $t = 3$), w.r.t γ; these EERs are obtained on the ImageCLEF set. The right-hand side figure, shows this EER class-by-class (when $\gamma = 1$) and a comparison against the underlying baseline kernel (i.e., $t = 0$); for ease of visualization we sort classes according to their difficulty (i.e., increasing error rates when using the baseline kernel). It is clear that our context-aware kernel (with $\gamma = 1$) decreases the EERs for almost all the classes.

$\mathbf{\Phi}_{\mathbf{x}}^{(0)}$ using the principal axes of KPCA[3].
According to Table (1) and Fig. (2), results obtained on different databases, clearly and consistently illustrate the out-performance of the learned context-aware kernel maps with respect to context-free ones for almost all the cases, with only few iterations ($t = 3$ in practice).

Kernels	Lin (CF)	Lin (CD)	Poly (CF)	Poly (CD)	RBF (CF)	RBF (CD)
Olivetti	1.50 ± 0.31	**0.28 ± 0.21**	1.44 ± 0.30	**0.26 ± 0.17**	1.36 ± 0.34	**0.66 ± 0.30**
Swedish	2.58 ± 0.90	**0.12 ± 0.11**	2.47 ± 0.97	**0.08 ± 0.01**	2.44 ± 0.85	**2.32 ± 0.94**
Kernels	Tri (CF)	Tri (CD)	χ^2 (CF)	χ^2 (CD)	HI (CF)	HI (CD)
Olivetti	1.00 ± 0.26	**0.30 ± 0.20**	1.50 ± 0.27	**0.28 ± 0.19**	0.90 ± 0.26	**0.28 ± 0.18**
Swedish	1.42 ± 0.72	**0.14 ± 0.13**	2.72 ± 0.91	**0.12 ± 0.12**	1.54 ± 0.77	**0.10 ± 0.01**

Table 1. This table shows EERs (in %) and standard deviations, of image annotation using SVMs, for different baseline kernels (denoted CF) and the underlying context-aware versions (denoted CD). In all these experiments, $N_a = N_r = 8$, the number of iterations = 3 and $\gamma = 1$.

4 Discussion

Similarity Diffusion. Our kernel is able to recursively diffuse the similarity from/to larger and more influencing contexts (i.e., two primitives are considered similar if their neighbors, with close spatial configurations, are similar and if the

[3] Note that KPCA approximation is exact when principal axes are learned using the whole set $\mathcal{X}$ [21] but computation is expensive, so we learn the principal axes on a small subset of $\mathcal{X}$.

neighbors of their neighbors are similar too, etc.) so resulting into a recursive definition and propagation/diffusion of similarity through the spatial structure of primitives (interest points in particular). Therefore, our context-aware kernel exploits pairwise (local) as well as higher order interactions (resulting from recursion). Our comparison in Table (1) corroborates this statement as the learned context-aware kernel maps show consistent gain compared to baseline kernel maps built upon shape context, SIFT and color histograms.

Invariance. It is easy to see that the adjacency matrices $\{\mathbf{P}_c\}$ are translation and rotation invariant and can also be made scale invariant when ϵ_p (see Eq. 1) is proportional to $\psi_s(\mathcal{S}_p)$. It follows that the context term of our kernel is invariant to 2D similarity transformations. Notice, also, that $\mathbf{S}$ in $\mathbf{K}^{(t)}$ involves similarity invariant (or at least tolerant) visual features $\psi_f(.)$ (including 128 SIFT features, color and shape context, see § 2), so both the kernels $\mathbf{K}^{(t)}$ and their explicit maps $\mathbf{\Phi}^{(t)}$ are similarity invariant.

Computational Complexity. let's consider a collection of N images including n interest points each. Assuming $\mathbf{\Phi}^{(t-1)}$, $\mathbf{K}^{(t-1)}$ known at iteration $t-1$, the worst complexity of evaluating the kernel map $\mathbf{\Phi}^{(t)}$ is $O(n^2N)$ and the complexity of the underlying SVM training using stochastic gradient descent is $O(N)$. This complexity reaches $O(n^4N^2)$ if one considers instead the evaluation of the gram matrix $\mathbf{K}^{(t)}$. As for testing, the complexity of evaluating the kernel map for a test image is $O(n^2)$ while the complexity of extending the gram matrix with that test image is $O(n^4N)$. Therefore, it becomes clear that the proposed kernel map evaluation method is at least an order of magnitude faster (than gram matrix based evaluation) both for training and testing.

If the adjacency matrices $\{\mathbf{P}_c\}$ are sparse, the method becomes even faster and its complexity reduces to $O(nM\ N)$, here $M = \max_{\mathbf{x}\in\mathcal{X}} |\mathcal{N}^c(\mathbf{x})| \ll n$. In practice, it takes less than 20 mins (on a standard 2Ghz PC) in order to evaluate the kernel maps and train the SVMs on the $N = 8,000$ pictures of ImageCLEF, instead of 5 hours when evaluating gram matrices before SVM learning.

5 Conclusion and take-home message

We introduced in this paper a kernel map learning procedure that takes into account the context. The purpose of this contribution is not to design another kernel; the main take home message is how to *upgrade* usual and widely used kernels, with context, in order to enhance their performances when used in SVM classification.

The proposed kernel design method shows a substantial gain compared to usual kernels for the challenging task of image classification. The method is also generic and could easily be extended to classification tasks in other neighboring fields and applications.

Acknowledgments

This work was supported in part by a grant from the Research Agency ANR (Agence Nationale de la Recherche) under the MLVIS project and a grant from DIGITEO under the RELIR project.

Appendix

Proof (of Proposition 1). Following (2), let us consider the function defined on the set of matrices in $\mathbb{R}^{m\times m}$

$$E : \mathbf{K} \mapsto -tr(\mathbf{K}\mathbf{S}') \ - \alpha \sum_c tr(\mathbf{K}\mathbf{P}_c\mathbf{K}'\mathbf{P}'_c) + \tfrac{\beta}{2}\left\|\mathbf{K}\right\|_2^2 \tag{5}$$

The necessary condition of the fixed-point relation in (4) results from $\partial E/\partial \mathbf{K} = 0$ (details about derivative are omitted in this proof). We will now prove that the function ψ is L-Lipschitzian, with $L = \gamma\| \sum_c \mathbf{P}_c \ \mathbf{1}_{mm} \ \mathbf{P}'_c\|_1$.
Given two matrices $\mathbf{K}^{(1)}$, $\mathbf{K}^{(2)}$, we have

$$\begin{aligned}
\left\|\mathbf{K}^{(2)} - \mathbf{K}^{(1)}\right\|_1 &= \sum_{\mathbf{x},\mathbf{x}'} \left|\mathbf{K}^{(2)}_{\mathbf{x},\mathbf{x}'} - \mathbf{K}^{(1)}_{\mathbf{x},\mathbf{x}'}\right| \\
&= \gamma \sum_{\mathbf{x},\mathbf{x}'} \left| \sum_{u,u',c} \mathbf{P}_{c,\mathbf{x},u} \ \mathbf{P}_{c,\mathbf{x}',u'} \ (\mathbf{K}^{(1)}_{u,u'} - \mathbf{K}^{(0)}_{u,u'}) \right| \\
&= \gamma \sum_{\mathbf{x},\mathbf{x}'} \left| \sum_{u,u'} (\mathbf{K}^{(1)}_{u,u'} - \mathbf{K}^{(0)}_{u,u'}) \sum_c \mathbf{P}_{c,\mathbf{x},u} \ \mathbf{P}_{c,\mathbf{x}',u'} \right| \\
&\leq \gamma \sum_{\mathbf{x},\mathbf{x}'} \sum_{u,u'} \left|\mathbf{K}^{(1)}_{u,u'} - \mathbf{K}^{(0)}_{u,u'}\right| \sum_c \left|\mathbf{P}_{c,\mathbf{x},u} \ \mathbf{P}_{c,\mathbf{x}',u'}\right| \\
&\leq \gamma \sum_{u,u'} \left|\mathbf{K}^{(1)}_{u,u'} - \mathbf{K}^{(0)}_{u,u'}\right| \times \sum_{\mathbf{x},\mathbf{x}'} \sum_{u,u',c} \left|\mathbf{P}_{c,\mathbf{x},u} \ \mathbf{P}_{c,\mathbf{x}',u'}\right| \\
&\quad \text{(as } \textstyle\sum_i |a_i|.|b_i| \leq \sum_{i,j} |a_i|.|b_j|, \ \forall \ \{a_i\}, \{b_j\} \subset \mathbb{R}) \\
&= L \ \left\|\mathbf{K}^{(1)} - \mathbf{K}^{(0)}\right\|_1
\end{aligned}$$

with $L = \gamma\left\| \sum_c \mathbf{P}_c \ \mathbf{1}_{mm} \ \mathbf{P}'_c\right\|_1$ □

Proof (of Proposition 2). We proceed by induction; $\mathbf{K}^{(0)}_{\mathbf{x},\mathbf{x}'} = \langle \psi_f(\mathbf{x}), \psi_f(\mathbf{x}')\rangle$ is explicit p.s.d as it is per definition p.s.d and the mapping $\psi_f(.)$ is known and finite dimensional. Assuming $\mathbf{K}^{(t)}_{\mathbf{x},\mathbf{x}'}$ explicit p.s.d, we obtain

$$\begin{aligned}
\mathbf{K}^{(t+1)}_{\mathbf{x},\mathbf{x}'} &= \left(\mathbf{\Phi}'^{(0)}\mathbf{\Phi}^{(0)} + \gamma \sum_c \mathbf{P}_c\mathbf{K}^{(t)}\mathbf{P}'_c\right)_{\mathbf{x},\mathbf{x}'} = \left(\mathbf{\Phi}'^{(0)}\mathbf{\Phi}^{(0)} + \gamma \sum_c \mathbf{P}_c\mathbf{\Phi}'^{(t)}\mathbf{\Phi}^{(t)}\mathbf{P}'_c\right)_{\mathbf{x},\mathbf{x}'} \\
&= \left(\mathbf{\Phi}'^{(t+1)}\mathbf{\Phi}^{(t+1)}\right)_{\mathbf{x},\mathbf{x}'}
\end{aligned} \tag{6}$$

$$\text{where} \quad \mathbf{\Phi}^{(t+1)} = \left(\mathbf{\Phi}'^{(0)} \quad \gamma^{\frac{1}{2}} \ \mathbf{P}_1\mathbf{\Phi}'^{(t)} \ \ldots \quad \gamma^{\frac{1}{2}}\mathbf{P}_{N_rN_a}\mathbf{\Phi}'^{(t)}\right)', \tag{7}$$

so $\mathbf{K}^{(t+1)}_{\mathbf{x},\mathbf{x}'}$ is also symmetric, continuous and p.s.d. Since $\mathbf{\Phi}^{(t)}$ is finite dimensional, $\mathbf{\Phi}^{(t+1)}$ defined in (7) is also finite dimensional so $\mathbf{K}^{(t+1)}_{\mathbf{x},\mathbf{x}'}$ is explicit p.s.d □

References

1. Bahlmann, C., Haasdonk, B., Burkhardt, H.: On-line handwriting recognition with support vector machines, a kernel approach. IWFHR, pages 49–54, (2002).
2. Barnard, K., Duygululu, P., Forsyth, D., Blei, D., Jordan, M.: Matching words and pictures. The Journal of Machine Learning Research, (2003).

3. Belongie, S., Malik, J., Puzicha, J.: Shape context: A new descriptor for shape matching and object recognition. NIPS, (2000).
4. Bottou, L.: Large scale machine learning with stochastic gradient descent. Proc of the 19th int conference on computational statistics, pages 177–187, (2010).
5. Boughorbel, S., Tarel, J., Boujemaa, N.: The intermediate matching kernel for image local features. IEEE International J. Conference on Neural Networks, (2005).
6. Carneiro, G., Vasconcelos, N.: Formulating semantic image annotation as a supervised learning problem. in Proc. of CVPR, (2005).
7. Gao, Y., Fan, J., Xue, X., Jain, R.: Automatic image annotation by incorporating feature hierarchy and boosting to scale up svm classifiers. ACM Multimedia, (2006).
8. Gartner, T.: A survey of kernels for structured data. Multi Relational Data Mining, 5(1):49–58, (2003).
9. Grauman, K., Darrell, T.: The pyramid match kernel: Efficient learning with sets of features. Journal of Machine Learning Research (JMLR), 8:725–760, (2007).
10. Haussler, D.: Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, U. of California in Santa Cruz, CS Department, July, (1999).
11. He, X., Zimel, R., Carreira, M.: Multiscale conditional random fields for image labeling. In CVPR, (2004).
12. Jaakkola, T., Diekhans, M., Haussler, D.: Using the Fisher kernel method to detect remote protein homologies. ISMB, pages 149–158, (1999).
13. Kondor, R., Jebara, T.: A kernel between sets of vectors. In proceedings of the 20th International conference on Machine Learning, (2003).
14. Li, J., Wang, J.Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. on PAMI, 25(9):1075–1088, (2003).
15. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2):91–110, (2004).
16. Lyu, S.: Mercer kernels for object recognition with local features. In the proceedings of the IEEE Computer Vision and Pattern Recognition, (2005).
17. Monay, F., GaticaPerez, D.: Plsa-based image autoannotation: Constraining the latent space. in Proc. of ACM International Conference on Multimedia, (2004).
18. Moreno, P., Ho, P., Vasconcelos, N.: A kullback-leibler divergence based kernel for svm classification in multimedia applications. NIPS, (2003).
19. Moser, G., Serpico, B.: Combining support vector machines and markov random fields in an integrated framework for contextual image classification. TGRS, (2012).
20. Nowak, S., Huiskes, M.: New strategies for image annotation: Overview of the photo annotation task at imageCLEF 2010. in Working Notes of CLEF 2010, (2010).
21. Scholkopf, B., Smola, A., Muller, K.-R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 10:1299–1319, (1998).
22. Semenovich, D., Sowmya, A.: Geometry aware local kernels for object recognition. In ACCV, (2010).
23. Shawe-Taylor, J., Cristianini, N.: Support vector machines and other kernel-based learning methods. Cambridge University Press, (2000).
24. Sahbi, H., Audibert, J.-Y., Keriven, R.: Context-Dependent Kernels for Object Classification, IEEE Trans on PAMI, Vol. 33, number. 4, April, (2011).
25. Singhal, A., Jiebo, L., Weiyu, Z.: Probabilistic spatial context models for scene content understanding. In CVPR, (2003).
26. Vapnik, V-N.: Statistical learning theory. A Wiley-Interscience Publication, 1998.
27. Wallraven, C., Caputo, B., Graf, A.: Recognition with local features: the kernel recipe. ICCV, pages 257–264, (2003).