Context-Based Support Vector Machines for Interconnected Image Annotation

Hichem Sahbi[1] and Xi Li[1,2,3]

[1] CNRS Telecom ParisTech, Paris, France
[2] School of Computer Science, The University of Adelaide, Australia
[3] NLPR, CASIA, Beijing, China
hichem.sahbi@telecom-paristech.fr, lixichinanlpr@gmail.com

Abstract. We introduce in this paper a novel image annotation approach based on support vector machines (SVMs) and a new class of kernels referred to as context-dependent. The method goes beyond the naive use of the intrinsic low level features (such as color, texture, shape, etc.) and context-free kernels, in order to design a kernel function applicable to interconnected databases such as social networks. The main contribution of our method includes (i) a variational approach which helps designing this function using both intrinsic features and the underlying contextual information resulting from different links and (ii) the proof of convergence of our kernel to a positive definite fixed-point, usable for SVM training and other kernel methods. When plugged in SVMs, our context-dependent kernel consistently improves the performance of image annotation, compared to context-free kernels, on hundreds of thousands of Flickr images.

1 Introduction

Recent years have witnessed a rapid increase of image sharing spaces, such as Flickr, due to the spread of digital cameras and mobile devices. An urgent need is how to effectively search these huge amounts of data and how to exploit the structure of these sharing spaces. A possible solution is CBIR (Content-Based Image Retrieval); where images are represented using low-level visual features (color, texture, shape, etc.) and searched by analyzing and comparing those features. However, low-level visual features are usually unable to deliver satisfactory semantics, resulting in a gap between them and the high-level human interpretations. To address this problem, a variety of machine learning techniques were introduced in order to discover the intrinsic correspondence between visual features and semantics of images and allow to predict keywords for images.

1.1 Related Work

Conventionally, image annotation is converted into a classification problem. Existing state of the art methods (for instance [1,2]) treat each keyword or concept

R. Kimmel, R. Klette, and A. Sugimoto (Eds.): ACCV 2010, Part I, LNCS 6492, pp. 214–227, 2011.

as an independent class, and then train the corresponding concept-specific classifier to identify images belonging to that class, using a variety of machine learning techniques such as hidden Markov models [2], latent Dirichlet allocation [3], probabilistic latent semantic analysis [4], and support vector machines [5]. The aforementioned annotation methods may also be categorized into two branches; region-based requiring a preliminary step of image segmentation [2,12], and holistic [6,25] operating directly on the whole image space. In both cases, training is achieved in order to learn how to attach keywords with the corresponding visual features.

The above annotation methods heavily rely on their visual features for image annotation. Due to the semantic gap, they are unable to fully explore the semantic information inside images. Another class of annotation methods has emerged that takes advantage of extra information (tags, context, users' feedback, ontologies, etc.) in order to capture the correlations between images and concepts. A representative work is the cross-media relevance model (CMRM) [6,9], which learns joint statistics of visual and concepts and its variants [7,8]. The model uses the keywords shared by similar images to annotate new ones. In [22], the similarity measure between images integrates contextual information for concept propagation. Semi-supervised annotation techniques were also studied and usually rely on graph inference [10,11,12,13]. The original work, in [3,26], is inspired from machine translation and considers images and keywords as two different languages; in that case, image annotation is achieved by translating visual words into keywords.

Other existing annotation methods focus on how to define an effective distance measure for exploring the semantic relationships between concepts in large scale databases. In [19], the Normalized Google similarity Distance (NGD) is proposed by exploring the textual information available on the web. It is a measure of semantic correlations derived from counts returned by Google's search engine for a given set of keywords. Following the idea of [19], the Flickr distance [20] is proposed to precisely characterize the visual relationships between concepts. Each one is represented by a visual language model in order to capture its underlying visual characteristics. Then, a Flickr distance is defined, between two concepts, as the square root of Jensen-Shannon (JS) divergence between the corresponding visual language models. Other techniques consider extra knowledge derived from ontologies (such as the popular WordNet [14,15,16]) in order to enrich annotations [21]. The method in [14] introduces a visual vocabulary in order to improve translation model in the preprocessing stage of visual feature extraction. A directed acyclic graph is used to model the causal strength between concepts, and image annotation is performed by inference on this graph [15]. In [17,18], the semantic ontology information is integrated in the post processing stage in order to further refine initial annotations.

1.2 Motivation and Contribution

Among the most successful annotation methods, those based on machine learning and mainly support vector machines; show a particular interest as they are

performant and theoretically well grounded [24]. Support vector machines [23], basically require the design of similarity measures, also referred to as *kernels*, which should provide high values when two images share similar structures/appearances and should be invariant, as much as possible, to the linear and non-linear transformations. They also satisfy positive definiteness which ensures, according to Vapnik's SVM theory [24], optimal generalization performance and also the uniqueness of the SVM solution. In practice, kernels should not depend only on intrinsic aspects of images (as images with the same semantic may have different visual and textual features), but also on different sources of knowledge including context.

In this paper, we introduce an image annotation approach based on a new family of kernels which take high values not only when images share the same visual content but also the same context. The context of an image is defined as the set of images, with the same tags, and exhibiting better semantic descriptions, compared to both pure visual and tag based descriptions. The issue of combining context and visual content for image retrieval is not new (see for instance [28,29,32]) but the novel part of this work aims to (i) integrate context, in kernel design useful for classification and annotation, and (ii) plug these kernels in support vector machines in order to take benefit from their well established generalization power [24]. This type of kernels will be referred to as context-dependent (CDK) while those relying only on the intrinsic visual or textual content will be referred to as context-free. Again, our proposed method goes beyond the naive use of low level features and context-free kernels (established as the standard baseline in image retrieval) in order to design a kernel applicable to annotation and suitable to integrate the "contextual" information taken from tagged links in interconnected datasets. In the proposed CDK, two images (even with different visual content and even sharing different tags) will be declared as similar if they share the same visual context[1]. This is usually useful as tags in interconnected data (such social networks) may be noisy and misspelled. Furthermore, the intrinsic visual content of images might not always be relevant especially for categories exhibiting large variation of the underlying visual aspects.

Through this work, an image database is modeled as a graph where nodes are pictures and edges correspond to the shared tagged links. We design our CDK as the fixed point of a constrained energy function mixing (i) a fidelity term which measures visual similarity between images, (ii) a context criterion that captures the similarity between the underlying links and (iii) a regularization term which helps defining a direct analytic solution. In the remainder of this paper we consider X as a random variable standing for all the possible images of the world, here X is drawn from an existing but unknown probability distribution P. Terminology and notation will be introduced as we go through different sections of this paper which is organized as follows: Section 2 tackles the issue

[1] For instance, two images I1, I2 (e.g., red Ferrari and black limousine) may be connected to two groups of "visually similar" images (two groups of cars are their contexts). If the two groups of images (groups of cars) are similar then one may conclude that the two images I1, I2 are also similar (see also Fig. 1 in [30]).

of kernel design, followed by some results about the positive definiteness and the convergence of the kernel to a fixed-point. Section 3 shows experimental results and the applicability of CDK in order to handle interconnected datasets including Flickr and NUSWIDE. We will conclude the method in Section 4 while providing promising extensions for a future work.

2 Kernel Design

Let us consider $\mathcal{X} = \{x_1, \ldots, x_n\}$ as a finite set of images drawn from the same distribution as X. Considering $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ as a continuous symmetric function which, given two images (x_i, x_j), provides us with a similarity measure; this function will be referred to as kernel. Our goal is to design $k(x_i, x_j)$ by taking into account the properties of x_i, x_j and also their links, i.e., the set of images which are connected to x_i, x_j.

2.1 Context and Graph-Links

We model an image database $\mathcal{X}$ using a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where nodes $\mathcal{V} = \{v_1, \ldots, v_n\}$ correspond to pairs $\{(x_i, \psi_f(x_i))\}_i$ and edges $\mathcal{E} = \{e_{i,j,\omega}\}$ are the set of tagged links of $\mathcal{G}$. In the above definition, $\psi_f(x_i)$ corresponds to the features of x_i (color, texture, shape, etc.) while $e_{i,j,\omega} = (v_i, v_j, \omega)$ defines a connection between v_i, v_j of type ω. The latter might be any particular tag for instance two images are linked when they share the same semantics, owners, GPS locations, etc. Through this work, two images are connected if they share the same Flickr-tag. *It is worth emphasizing that these tags are different from the concepts (classes) used for training and annotation. Indeed, each image belongs to one or multiple concepts which are different from the tags provided by users (see Section 3).*

Now, introduce the context (or neighborhood) $\mathcal{N}^\omega(x_i) = \left\{x_j : (x_i, x_j, \omega) \in \mathcal{E}\right\}$. This definition of context $\{\mathcal{N}^w(x)\}_w$ reflects the co-occurrence of different images with particular connection types (again defined using tags).

2.2 Context-Dependent Kernel Design

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

$$\min_{\substack{\mathbf{K} \geq 0 \\ \|\mathbf{K}\|_1 = 1}} \quad \mathrm{Tr}\big(\mathbf{K}\ \mathbf{D}^{'}\big) + \beta\ \mathrm{Tr}\big(\mathbf{K}\ \log\ \mathbf{K}^{'}\big)\ -\ \alpha \sum_{\omega} \mathrm{Tr}\big(\mathbf{K}\ \mathbf{P}_\omega\ \mathbf{K}'\ \mathbf{P}'_\omega\big)$$

Here 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 $\mathrm{Tr}(\cdot)$ denotes matrix trace. The non matrix form of the above objective function may be written $\sum_{x,x'} k(x,x')d(x,x') + \beta \sum_{x,x'} k(x,x')\log(k(x,x')) - \alpha \sum_{\omega,x,x',y,y'} k(x,x')k(y,y')$ (with $y \in \mathcal{N}^\omega(x), y' \in \mathcal{N}^\omega(x')$). The first term, in the above constrained minimization problem, measures the quality of matching two feature vectors $\psi_f(x)$, $\psi_f(x')$. In the case of visual features, this is considered as the distance, $d(x,x')$, between the visual descriptors (color, texture, shape, etc.) 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 visual features, the probability distribution $\{k(x,x')\}$ should be flat so the negative of the entropy is minimized. This term also helps to define a simple solution and solve the constrained minimization problem easily. The third term is a context criterion which considers that a high value of $k(x,x')$ should imply high kernel values in the respective neighborhoods $\mathcal{N}^\omega(x)$ and $\mathcal{N}^\omega(x')$ of x and x'. We formulate the minimization problem by adding an equality constraint and bounds which ensure a normalization of the kernel values and allow to see $\mathbf{K}$ as a joint probability distribution (or P-Kernel [33]).

2.3 Solution

The above optimization problem admits a solution $\tilde{\mathbf{K}}$, which is the limit of the context-dependent kernels $\mathbf{K}^{(t)} = G(\mathbf{K}^{(t-1)})/\|G(\mathbf{K}^{(t-1)})\|_1$, with $G(\mathbf{K}) = \exp\left\{-\frac{\mathbf{D}}{\beta} + \frac{\alpha}{\beta}\sum_\omega \left(\mathbf{P}_\omega \mathbf{K} \mathbf{P}'_\omega + \mathbf{P}'_\omega \mathbf{K} \mathbf{P}_\omega\right)\right\}$, and $\mathbf{K}^{(0)} = \exp(-\mathbf{D}/\beta)/\|\exp(-\mathbf{D}/\beta)\|_1$. By taking small enough α, convergence of this kernel to a fixed point is satisfied (see [30]). 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 images. Besides our choice of $\mathbf{K}^{(0)}$ is exactly the optimum (and fixed point) for $\alpha = 0$. Detailed proof of this solution and its convergence to a fixed point may be found in [30].

2.4 Positive Definiteness

A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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 ϕ 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 1. *The context-dependent kernels on $\mathcal{X}$ defined in (2.3) by the matrices $\tilde{\mathbf{K}}$ and $\mathbf{K}^{(t)}$, $t \geq 0$, are positive definite.*

Proof. See [30]. □

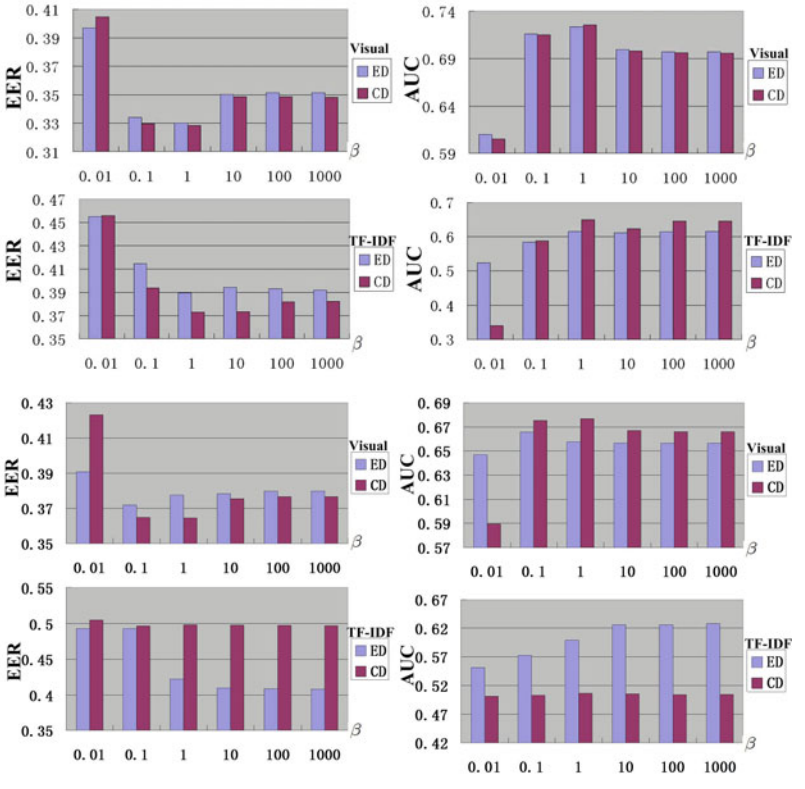

Fig. 1. This figure shows the annotation performances of different βs under different metrics on the MIRFLICKR-25000 (two first rows) and NUSWIDE datasets (other rows)

3 Benchmarking

This section evaluates the performance of image annotation tasks and shows the extra advantage of our context-dependent kernel (CDK) with respect to the use of many existing context-free ones such as the gaussian, the polynomial, the chi-square, etc. The point here is also to show the importance of the context in kernel design through different databases and settings.

3.1 Databases and Settings

We evaluated CDK on the MIRFLICKR-25000[2] as well as the NUSWIDE[3] datasets. Both sets are challenging; the first one, MIRFLICKR-25000 contains 25,000 images belonging to 24 concepts (for instance "sky, clouds, water, sea,

[2] http://press.liacs.nl/mirflickr/
[3] http://lms.comp.nus.edu.sg/research/NUSWIDE.htm

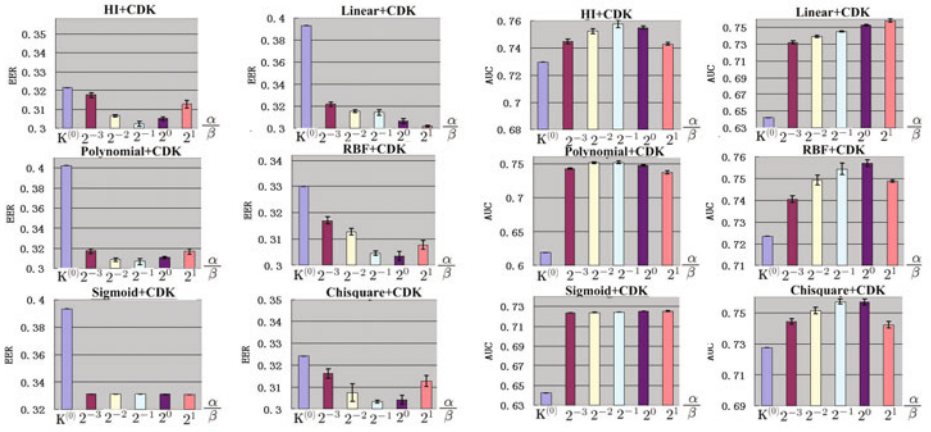

Fig. 2. This figure shows the performances of annotation on the MIRFLICKR-25000 dataset (with $\beta = 1$). It includes EER and AUC annotation performances of six context-dependent kernels based on visual features. Compared with the underlying baseline kernels, the six (best) context-dependent kernels achieve a relative gain of respectively 5.91%, 23.09%, 23.69%, 8.03%, 17.34%, and 6.47% for EER. Means and standard deviations are taken over 20 trials.

river,...") while the second one, NUSWIDE database, is larger and contains more than a quarter of a million of images (exactly 269,648) belonging to 81 concepts. Note that both sets were downloaded from Flickr through its public API.

Each image in MIRFLICKR-25000 is processed in order to extract the bag-of-word SIFT representation [27]. Precisely, SIFT features are extracted at three different spatial pyramid levels, and quantized into 200 codewords. Consequently, the visual feature for each image is a 4200-dimensional concatenated histogram of three spatial pyramid levels. Moreover, images in MIRFLICKR-25000 are supplied with tags (which again are different from the concepts used for learning and annotation, see Section 2.1). In total, 1,386 tags are used, each one annotates at least 20 images. As a matter of comparison, textual features are also used. Indeed, each image characterized by its tags, is mapped using the TF-IDF (term frequency-inverse document frequency) resulting into a feature vector of 1,386 dimensions. Images in the NUSWIDE set are also indexed with the bag-of-word SIFT features of 500 dimensions and they are also supplied with 1,000 tags used in order to extract the TF-IDF features.

Let Ω denote the union of tags over all the images of a given set (either MIRFLICKR-25000 or NUSWIDE). Again, we define the underlying graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, here nodes $\mathcal{V}$ are defined similarly as in Section 2.1, whereas edges are defined as $\mathcal{E} = \{e_{i,j,\omega} : \omega \in \Omega, \#\omega \in [T_D, T_U]\}$. Here $\#\omega$ denotes the number of images tagged by ω and T_D, T_U are two fixed thresholds; their setting determines the complexity and topology of the graph and may also affect performance as shown later in this section.

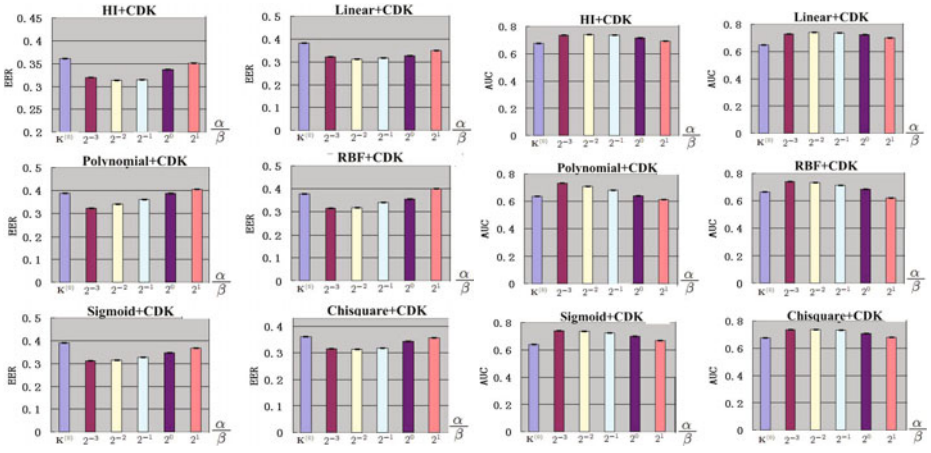

Fig. 3. This figure shows the performances of annotation on the NUSWIDE dataset. It includes EER and AUC annotation performances of six context-dependent kernels based on visual features (with $\beta = 1$). Compared with the underlying baseline kernels, the six (best) context-dependent kernels achieve a relative gain of respectively 12.98%, 18.28%, 12.76%, 16.18%, 19.23%, and 12.94% for EER. Means and standard deviations are taken over 20 trials.

3.2 Hold-Out Generalization and Comparison

We evaluate $\mathbf{K}^{(t)},\ t \in \mathbb{N}^+$ using six power assist settings

(i) Linear $\quad \mathbf{K}^{(0)}_{x,x'} = \langle \psi_f(x), \psi_f(x') \rangle$,

(ii) Polynomial $\quad \mathbf{K}^{(0)}_{x,x'} = (\langle \psi_f(x), \psi_f(x') \rangle + 1)^2$,

(iii) RBF $\quad \mathbf{K}^{(0)}_{x,x'} = \exp(-\|\psi_f(x) - \psi_f(x')\|_2/\beta)$,

(iv) Histogram intersection $\quad \mathbf{K}^{(0)}_{x,x'} = \sum_i \min(\psi_f(x)_i, \psi_f(x')_i)$,

(v) Chisquare $\quad \mathbf{K}^{(0)}_{x,x'} = 1 - \frac{1}{2} \sum_i \frac{(\psi_f(x)_i - \psi_f(x')_i)^2}{(\psi_f(x)_i + \psi_f(x')_i)}$,

(vi) Sigmoid $\quad \mathbf{K}^{(0)}_{x,x'} = \tanh(\ \langle \psi_f(x), \psi_f(x') \rangle)$.

These context-free kernels are plugged into the underlying CDK kernels $\mathbf{K}^{(t)},\ t \in \mathbb{N}^+$, and the resulting CDKs will be referred to as "Linear+CDK", "Poly+CDK", "RBF+CDK", "HI+CDK", "Chisquare+CDK", and "Sigmoid+CDK" respectively. 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 "one-versus-all" SVM classifiers[4] for each concept in

[4] http://www.csie.ntu.edu.tw/∼cjlin/libsvm

the MIRFLICKR-25000 and the NUSWIDE datasets. For each concept, training is achieved using three-random folds ($\sim 75\%$) of the data while testing is achieved on the remaining-fold. Notice that this process is randomized 20 times and the outputs of the underlying SVM classifiers are taken as the average values through these 20 random samplings; this makes classification results less sensitive to sampling and unbalanced classes.

Evaluation measures. Performances are reported, on different test sets, using the hold-out equal error rate (EER) and area under ROC (receiver operating characteristics) curve (AUC). The EER is the balanced generalization error which equally weights the positive and the negative errors. It can be easily computed from the ROC curve. The smaller the EER, the better the annotation performance. The AUC measures the ranking quality of a classifier, and it can be viewed as an estimation of the probability that the classifier ranks a randomly selected positive sample higher than a randomly selected negative sample. The larger the AUC, the better the annotation performance. These two measures are evaluated using the standard script provided by the ImageClef evaluation campaigns.

Context-free kernel setting. We aim to explore the optimal parameter settings for β under an appropriate dissimilarity metric $d(x, x')$. Thus, two popular metrics are used for performance evaluations: (i) Euclidean distance (ED); and (ii) Chisquare distance (CD).

$$\begin{aligned} &\text{(i)} \quad \text{ED} \quad d(x,x') = \|\psi_f(x) - \psi_f(x')\|_2, \\ &\text{(ii)} \quad \text{CD} \quad d(x,x') = \frac{1}{2}\sum_i \frac{(\psi_f(x)_i - \psi_f(x')_i)^2}{(\psi_f(x)_i + \psi_f(x')_i)}. \end{aligned}$$

Based on the above two metrics, we tuned the scale parameter β to achieve the best EER and AUC annotation performances. Fig. 1 shows the EER and AUC annotation results of different βs under different metrics using two features on the MIRFLICKR-25000 and the NUSWIDE datasets. Clearly, we found that the best performances are achieved on the MIRFLICKR-25000 dataset when $\beta = 1$ using the CD metric for both visual and TF-IDF features. As for NUSWIDE, it performs best when $\beta = 1$ (resp. $\beta = 1000$) using the CD (resp. ED) metric for visual (resp. TF-IDF) features.

Influence of the context. All the reported results show that the influence of the right-hand side of $\mathbf{K}^{(t)}$, $\alpha \neq 0$ increases as α increases (see Fig. 2), nevertheless and as shown in [30], the convergence of $\mathbf{K}^{(t)}$ to a fixed point is guaranteed only if α is bounded. When convergence is not guaranteed, CDK may suffer numerical instabilities resulting into degeneracy of the performance. Therefore it is obvious that α should be set to the highest possible value which also satisfies an upper bound criterion (see [30]). In these diagrams, the weight α is taken from five different values using a logarithmic scale $2^{-3}\beta$, $2^{-2}\beta$, $2^{-1}\beta$, $2^{0}\beta$, and $2^{1}\beta$.

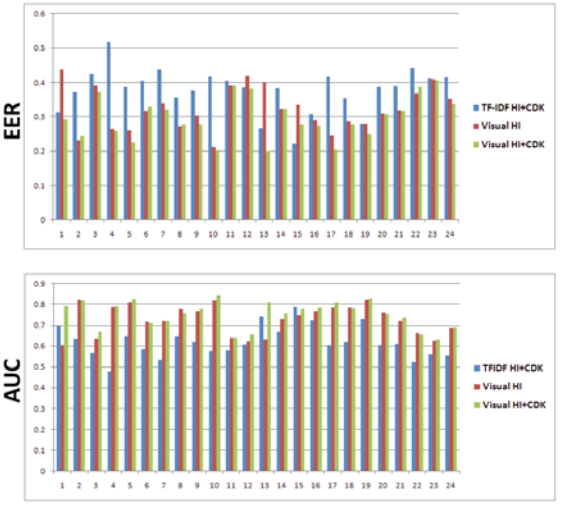

Fig. 4. This figure shows the performances of annotation, concept-by-concept, on the MIRFLICKR-25000 set. It includes EER (top) and AUC (bottom) annotation performances of the (best) baseline context-free kernel (**HI**), and the underlying context-dependent kernel (**HI+CDK**). Comparison is also shown with respect to the (best) context-dependent kernel based on TF-IDF (**TF-IDF HI+CDK**). Compared with **TF-IDF HI+CDK** (resp. visual **HI**), the average relative gain of visual **HI+CDK** is 19.83% (resp. 7.20%) for EER and 22.09% (resp. 3.64%) for AUC. The x-axis values correspond to the concept indices (in the same order as the one given in the original database).

Comparison. Fig. 2 shows the annotation performance (EER and AUC) of the context-dependent kernel using the six power assist settings defined earlier. Specifically, global annotation performances using bag-of-word visual feature are plotted in Fig. 2 for MIRFLICKR-25000 and Fig. 3 for NUSWIDE, where the x-axis of each sub-figure corresponds to different settings of α and the y-axis shows the underlying error rates ($\mathbf{K}^{(0)}$ corresponds to the baseline context-free kernels). Correspondingly, concept-by-concept annotation performances using bag-of-word visual feature and TF-IDF textual feature are plotted in Fig. 4 for MIRFLICKR-25000 and Fig. 5 for NUSWIDE. According to Fig. 2, performances of context-dependent kernels are mostly better than those of context-free ones, with just two iterations (i.e., $t \geq 2$). From Figs. 4-5, it is seen that the average concept-by-concept EER (resp. AUC) performances for the visual feature are mostly lower (resp. higher) than those for the TF-IDF textual feature. Standard deviations are also shown with respect to different threshold intervals $[T_D, T_U]$ in Fig. 2. This clearly corroborates the statement that improvement is not only due to the nature of the features but also to the integration of the context in kernel design.

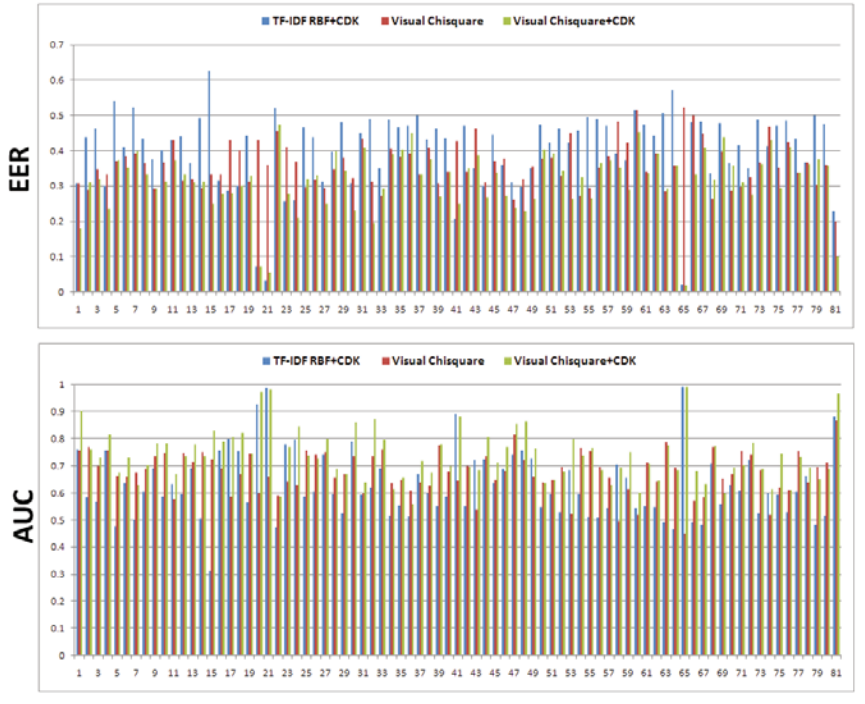

Fig. 5. This figure shows the performances of annotation, concept-by-concept, on the NUSWIDE dataset. It includes EER (top) and AUC (bottom) annotation performances of the (best) baseline context-free kernel (**Chisquare**), and the underlying context-dependent kernel (**Chisquare+CDK**). Comparison is also shown with respect to the (best) context-dependent kernel based on TF-IDF (**TF-IDF RBF+CDK**). Compared with **TF-IDF RBF+CDK** (resp. **Chisquare**), the average relative gain of **Chisquare+CDK** is 19.43% (resp. 11.34%) for EER and 19.94% (resp. 10.32%) for AUC. The x-axis values correspond to the concept indices (in the same order as the one given in the original database).

Further illustrations, taken from MIRFLICKR-25000 database, show annotation results of the best context-dependent kernel **HI+CDK** and comparison with respect to the underlying ground truth annotation. The final annotation results of these twelve images are shown in Fig. 6. It is clear that our proposed method achieves reasonable annotation results.

Runtime. Note that training time depends on concepts and training set cardinalities. For instance the computation of a 8000^2 sub-block of the gram matrix $\mathbf{K}^{(t)}$ requires about 7 minutes on a standard 2.6GHZ PC with 2G memory. This time scales linearly w.r.t the number of sub-blocks. Assuming $\mathbf{K}^{(t-1)}$ known for a given pair x, x', the worst complexity of evaluating $\mathbf{K}^{(t)}$ is $O(\max(N^2, s))$, where s is the dimension of $\psi_f(x)$ and $N = \max_{x,\omega} \#\{\mathcal{N}^\omega(x)\}$. When $N < \sqrt{s}$,

Fig. 6. This table shows comparisons of ground truth and the best context-dependent kernel (**HI+CDK**) annotations on twelve images from MIRFLICKR-25000

the complexity of evaluating the proposed kernel is strictly equivalent to that of usual kernels such as the linear (N may be forced to a small value by sampling context).

4 Conclusion

We introduced in this work a novel approach for kernel design dedicated to interconnected datasets including social networks. The strength of this method resides in the inclusion of context links in kernel design thereby improving annotation performances consistently.

The "Take Home Message" is to show that the information present into a picture can be described *not only* by its intrinsic visual features (suffering the semantic gap) but also by the set of images in its "context". The proposed kernel gathers many fundamental properties (i) second order context criterion which

captures links between images (ii) well motivated definition of kernels via an energy function ending with a probabilistic interpretation.

Extensions of this work include the use of ontologies in order to enrich social link types. Other future work will exploit the positive definiteness of CDK in order to use lossless acceleration techniques suitable for even larger scale networks.

Acknowledgement. This work is supported by the French National Research Agency (ANR) under the AVEIR project.

References

1. Carneiro, G., Vasconcelos, N.: Formulating semantic image annotation as a supervised learning problem. In: Proc. of CVPR (2005)
2. Li, J., Wang, J.Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. on PAMI 25(9), 1075–1088 (2003)
3. Barnard, K., Duygululu, P., Forsyth, D., Blei, D., Jordan, M.: Matching words and pictures. The Journal of Machine Learning Research (2003)
4. Monay, F., GaticaPerez, D.: PLSA-based Image AutoAnnotation: Constraining the Latent Space. In: Proc. of ACM International Conference on Multimedia (2004)
5. Gao, Y., Fan, J., Xue, X., Jain, R.: Automatic Image Annotation by Incorporating Feature Hierarchy and Boosting to Scale up SVM Classifiers. In: Proc. of ACM MULTIMEDIA (2006)
6. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proc. of ACM SIGIR, pp. 119–126 (2003)
7. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: Proc. of NIPS (2004)
8. Feng, S., Manmatha, R., Lavrenko, V.: Multiple Bernoulli relevance models for image and video annotation. In: Proc. of ICCV, pp. 1002–1009 (2004)
9. Liu, J., Wang, B., Li, M., Li, Z., Ma, W., Lu, H., Ma, S.: Dual cross-media relevance model for image annotation. In: Proc. of ACM MULTIMEDIA, pp. 605–614 (2007)
10. Wan, X., Yang, J., Xiao, J.: Manifold-ranking based topic-focused multi-document summarization. In: Proc. of IJCAI, pp. 2903–2908 (2007)
11. Zhou, D., Weston, J., Gretton, A., Bousquet, O., Schölkopf, B.: Ranking on data manifolds. In: Proc. of NIPS (2004)
12. Liu, J., Li, M., Liu, Q., Lu, H., Ma, S.: Image annotation via graph learning. Pattern Recognition 42(2), 218–228 (2009)
13. Liu, J., Li, M., Ma, W., Liu, Q., Lu, H.: An adaptive graph model for automatic image annotation. In: Proc. of ACM International Workshop on Multimedia Information Retrieval, pp. 61–70 (2006)
14. Srikanth, M., Varner, J., Bowden, M., Moldovan, D.: Exploiting ontologies for automatic image annotation. In: Proc. of SIGIR, pp. 552–558 (2005)
15. Wu, Y., Chang, E.Y., Tseng, B.L.: Multimodal metadata fusion using causal strength. In: Proc. of ACM MULTIMEDIA, pp. 872–881 (2005)
16. Miller, G.A.: Wordnet: a lexical database for English. ACM Commun. 38(11), 39–41 (1995)
17. Wang, C., Jing, F., Zhang, L., Zhang, H.J.: Image annotation refinement using random walk with restarts. In: Proc. of ACM MULTIMEDIA, pp. 647–650 (2006)

18. Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple evidence & wordNet. In: Proc. of ACM MULTIMEDIA, pp. 706–715 (2005)
19. Cilibrasi, R., Vitanyi, P.M.B.: The google similarity distance. IEEE Transactions on Knowledge and Data Engineering (2007)
20. Wu, L., Hua, X., Yu, N., Ma, W., Li, S.: Flickr distance. In: Proc. of ACM MULTIMEDIA (2008)
21. Wang, Y., Gong, S.: Translating Topics to Words for Image Annotation. In: Proc. of ACM CIKM (2007)
22. Lu, Z., Ip, H.H.S., He, Q.: Context-Based Multi-Label Image Annotation. In: Proc. of ACM CIVR (2009)
23. Boser, B., Guyon, I., Vapnik, V.: An training algorithm for optimal margin classifiers. In: Fifth Annual ACM Workshop on Computational Learning Theory, Pittsburgh (1992)
24. Vapnik, V.: Statistical Learning Theory. A Wiley-Interscience Publication, Hoboken (1998)
25. Wang, C., Yan, S., Zhang, L., Zhang, H.: Multi-Label Sparse Coding for Automatic Image Annotation. In: Proc. of CVPR (2009)
26. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
27. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: Proc. of CVPR (2006)
28. Gallagher, A.C., Neustaedter, C.G., Cao, L., Luo, J., Chen, T.: Image Annotation Using Personal Calendars as Context. In: Proc. of ACM Multimedia (2008)
29. Cao, L., Luo, J., Huang, T.S.: Annotating Photo Collection by Label Propagation According to Multiple Similarity Cues. In: Proc. of ACM Multimedia (2008)
30. Sahbi, H., Audibert, J.-Y.: Social network kernels for image ranking and retrieval. In Technical Report, N 2009D009, TELECOM ParisTech (March 2009)
31. Shawe-Taylor, J., Cristianini, N.: Support Vector Machines and other kernel-based learning methods. Cambridge University Press, Cambridge (2000)
32. Yang, Y.H., Wu, P.T., Lee, C.W., Lin, K.H., Hsu, W.H., Chen, H.: ContextSeer: Context Search and Recommendation at Query Time for Shared Consumer Photos. In: Proc. of ACM Multimedia (2008)
33. Haussler, D.: Convolution Kernels on Discrete Structures. In Technical Report UCSC-CRL-99-10, University of California in Santa Cruz, Computer Science Department (July 1999)