Social network kernels for image ranking and retrieval

Noyaux de similarité pour les réseaux sociaux

Hichem Sahbi
Jean-Yves Audibert

2009D009

Mars 2009

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

Social Network Kernels for Image Ranking and Retrieval

Noyaux de Similarité pour les Réseaux Sociaux

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

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

Abstract

The exponential growth of social networks currently makes them the standard way to share and explore data where users put informations (images, text, audio,...) and refer to other contents (profiles, images,...). This creates connected networks whose links provide valuable informations in order to enhance the performance of many tasks in information retrieval including ranking and annotation.

We introduce in this paper a novel image retrieval framework based on a new class of kernels referred to as "network or context-dependent". The contribution of our method includes (i) a variational framework which helps designing a similarity (and ranking) using both the intrinsic image attributes and the underlying contextual informations resulting from different (e.g. social) links and (ii) the proof of convergence of the similarity function to a fixed-point. We will also show that the resulting fixed-point defines indeed a Mercer kernel in some reproducing kernel Hilbert space (RKHS). Experiments conducted on social network data mainly Flickr show the outperformance and the substantial gain of our ranking scheme with respect to the use of classic "context-free" similarity.

Résumé

La croissance exponentielle des réseaux sociaux sur Internet les rend actuellement des standards incontournables de partage et d'exploration des données multimédia. Dans ces réseaux, les utilisateurs rajoutent des informations (images, texte, audio,...) et créent des liens vers d'autres contenus. Ces liens sociaux fournissent des statistiques sur les données et permettent aussi d'améliorer les performances de plusieurs taches en recherche d'information comme le "ranking" et l'enrichissement des annotations.

On introduit dans cet article une nouvelle approche de recherche d'images basée sur une famille de noyaux dite dependente des "réseaux sociaux". La contribution de ce travail inclut (i) une approche variationnelle permettant de construire ces noyaux de similarité (et de "ranking") en utilisant les attributs visuels intrinsèques des images ainsi que leur contexte issue des liens sociaux et (ii) une preuve de convergence du noyau construit vers un point-fixe. On démontre aussi que ce dernier définit bien un espace à noyau reproduisant (RKHS). Les experiences, menées sur des données du réseau social *Flickr*, démontrent clairement les bonnes performances de notre noyau de "ranking" par rapport aux fonctions de similarité classiques c.a.d. indépendantes du contexte.

Keywords: Kernel Methods, Kernel Design, Context Based and Graph Similarity, Social Networks, Cross-Media Retrieval.

1. Introduction

Social networks (SN) such as Facebook, Flickr, MySpace's and Google's FriendConnect, are becoming trendy information sharing spaces which allow Web users to put and refer other contents (ECIR, 2009; WSIRTEL, 2007). For instance, personnel pictures in Flickr are meshed together when they share common "semantics", owners or interests (see Fig 1, B) implying a form of implicit social structure. This networking scheme is currently a valuable source of statistics both for real-world applications[1] and from the methodological point-of-view as links between images may boost ranking and retrieval performances (Russell et al., 2008).

Most of the current image querying paradigms (Ritendra et al., 2008) are based on ranking strategies which are clearly not appropriate in order to handle SN image data. This comes from the well known statistical inconsistency of the underlying low level features with respect to the user's "class of interest", the impossibility of these paradigms to quantify the user's "subjectivity" related to his social links and also complexity of images. More suitable techniques started to emerge for SN information retrieval mainly in closely related area such as text document ranking (Kirchhoff et al., 2008; Zhou et al., 2008; Hoser, 2009; Bian et al., 2008) and tagging (Ames and Naaman, 2007; Franke et al., 2007). Recent, but very few work, is now handling SN *image* retrieval such as the pioneering paper of Li et al. (2008) that uses visual links in order to propagate image tags and the method of Stone et al. (2008) which is of a particular interest; as authors consider face recognition in SN using conditional random fields and show a significant improvement when using contextual-links between faces in the recognition process. Other work uses graph transduction (Wang et al., 2008) and also manifold learning (Hoi et al., 2008; Sahbi et al., 2008b) in order to exploit links between data and performs metric learning/ranking. Manifold learning is a suitable technique in order to define distances relying on the topology of data where the general assumption states that locally distances are Euclidean but globally they are geodesics (Roweis and Saul, 2000). This strong assumption, which works reasonably well for many applications, is not well adapted in order to handle social network data as continuity with respect to the underlying links is not guaranteed.

Our goal in this work is to design a new family of kernels which take high values *not only when images share the same intrinsic visual features but also the same context or SN links.* The word kernel is used in order to designate similarity and also ranking metrics as one may define the latter from RKHS kernels and vice-versa (see Section 2.4). In the remainder of this paper, kernels based on intrinsic visual features will be referred to as

[1] FaceBook played a major role in the 2008 US presidential election and continue to be a major actor in profile-based publicity.

”network or context-free” while those including the outward (SN) contextual links will be named “network or context dependent”.

1.1 Background and Motivations

Context has played an important role in computer vision and mainly in scene interpretation where couple of context-dependent recognition and retrieval techniques show their out-performance with respect to context-free ones (Amores et al., 2005; Sahbi et al., 2008a; Jin and Geman, 2006; Bach, 2008; Zhu and Mumford, 2004; Geman and Johnson, 2002). The general idea is to capture the visual appearance of “objects of interest” in query images as well as the underlying visual context. This was initially motivated by the success of other closely related areas in automatic speech recognition and machine translation (see for instance Koehn et al. (2003)). Context based scene interpretation and retrieval consists in segmenting and extracting objects of interest in queries and modeling their contextual and spatial relationships (Galleguillos et al., 2008), then matching them with other images (Sahbi et al., 2008a). Regardless, the difficulty of segmenting and recognizing objects, relying only on the visual information is known to be limited by the semantic gap (Smeulders et al., 2000).

If one handles social network data, such as Flickr, then tags can be used in order to partially close the semantic gap. Considering the image query in the right-center of figure (1, B), the latter shares with other images a finite number of labels (“race car”, “car”, “Malibu”). If one is interested in finding cars in the same visual context as that query, and if one *conditions* the search using a conjunction (resp. disjunction) of those labels then it will not be possible to retrieve cars out of Malibu (resp. possible to find Malibu but not cars) resulting into bad recall/precision performances. Needless to say, visual and text informations are complementary in order to improve both recall and precision as already mentioned in many studies in cross-media retrieval (see for instance Arni et al. (2008)) and as will be shown through this paper but for the particular case of social networks.

1.2 Goals and Contributions

In this paper we introduce a ranking method based on a new family of kernels referred to as “social-network-kernels” (SNK). An image database is modeled as a graph where nodes are pictures and edges correspond to the social links. We design our SNK as the fixed point of a constrained energy function mixing (i) a fidelity term which measures intrinsic visual similarity between images, (ii) a neighborhood criterion that captures the context, i.e., resemblance between the underlying social links and (iii) a regularization term which helps finding a smooth solution in the form of a probability distribution. Notice that this work is different from the one in Sahbi et al. (2008a) in many aspects including

(i) The update of our objective function (see equation 1) which now considers kernel similarity between images and not interest points. Indeed, in our previous work, kernel design was achieved for interest points in the context of their images, while in the current version, kernel building is done for images in the context of their visual parts. One of the *key aspects* of this work resides in the new definition of these parts which are actually not explicitly

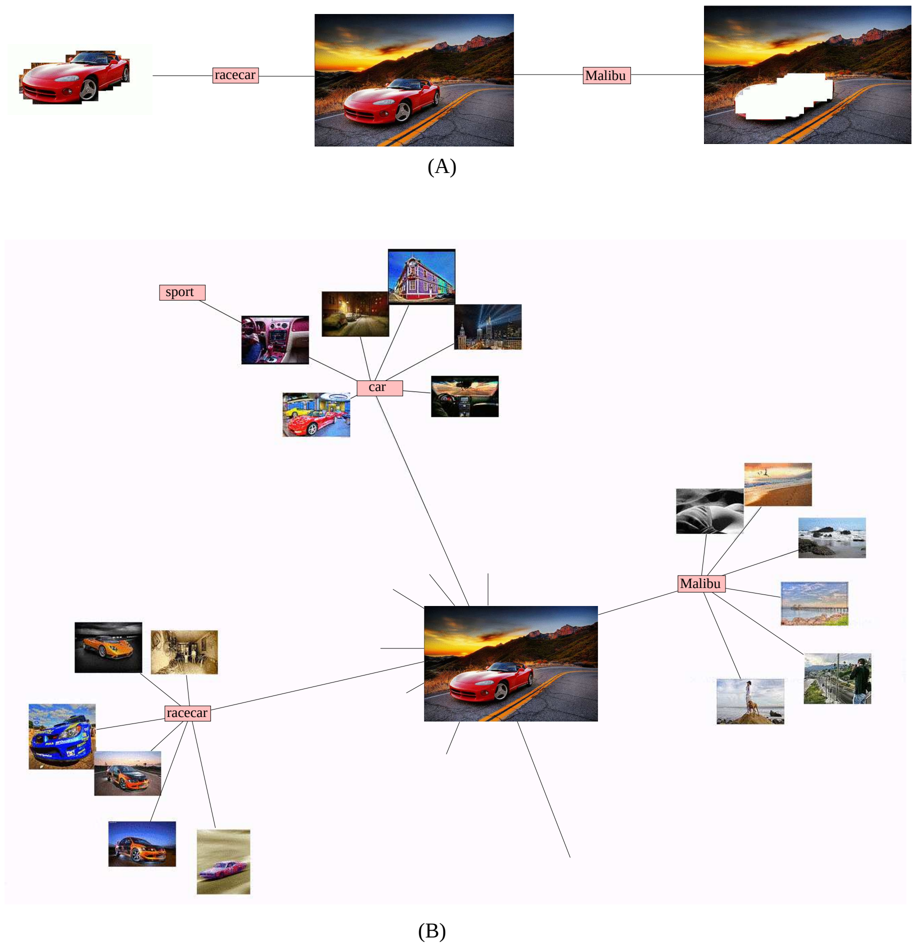

Figure 1: The image in the right-center in both (A) and (B) has many links to other pictures which share similar concepts, such as location or car style. The visual context of that query contains at least three objects “car”, “road” and “Malibu landscape”. While visually these objects are difficult to extract and to describe (A), they can be characterized by three classes of social links (B).

extracted (using segmentation or object extraction techniques) but instead *defined by the set of outward links pointing other images in the social network* (see Fig 1, B). This not only avoids us solving ill posed and difficult object segmentation problems (see Fig 1, A) but also provides us with more statistics about contextual-parts of images through their external social links.

(ii) The structure of social networks defines around each query a *bag-of-context images*; as will be shown in section (2) our approach is able to map these “order-less” and “variable-length” bags into fixed length and ordered feature vectors. This follows from the positive definiteness of SNK, i.e., there exists a reproducing Hilbert space where our kernel is seen as a dot product (see Section 2.4).

(iii) We will rise other theoretical issues mainly the convergence of our kernel to a fixed-point (see Section 2.3) and we will also show that, under particular conditions, one may derive a metric from SNK which is equivalent to the diffusion map distance (Lafon et al., 2006).

Globally, SNK captures intrinsic visual similarity as well as social network topology since its general form considers the similarity between any two images by incorporating also their context, i.e., the similarity of the surrounding bags of images in the social links. Our kernel can be viewed as a variant of "dynamic programming" similarity (Bahlmann et al., 2002) where instead of using the ordering criterion we consider a neighborhood assumption which states that two images are very similar if they have similar visual features and if they satisfy a neighborhood criterion, i.e., their SN linked images are similar too. Our kernel design might be seen as a variant of existing Markov based methods such as Fisher kernels (Jaakkola et al., 1999), which implement the conditional dependency between data. SNK also implements such dependency with an extra advantage of handling the context at different orders[2] and being the fixed point and the (sub)optimal solution of a constrained energy function closely related to the goal of our application, i.e., gathering (or at least balancing) discrimination and flexibility when taking into account the context. Again, one may use a simpler way in order to integrate the context by concatenating bags of neighboring images but these semi-structured data have no-fixed length and cannot be ordered so the framework presented later in the paper will map them into a dot product space (see Section 2.4). Finally, through all the sections of this paper, the word social network will be abusively and repeatedly used but one should not consider this work applicable only for social networks, as it can be naturally extended to other networked datasets (web-pages, etc.) and also to other pattern recognition tasks including classification and regression. Nevertheless, social networks are particular datasets which offer lots of informations (including labels and tags) that are currently well spread and easy to get on the Web.

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 of social network kernel design, followed by proofs that this function is indeed a Mercer kernel and convergent to a fixed-point. Section (3) shows experimental results and the applicability of SNK in order to handle SN databases including Flickr. We will discuss the method in Section (4) and conclude in (5) while providing some extensions for a future work.

2. Network-Dependent Similarity

Let us consider $\mathcal{X} = \{x_1, \dots, x_n\}$ as a finite set of images drawn from the same distribution as X. Considering $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as a continuous symmetric function which, given two

[2] As will be discussed later in the paper, the definition of the context is recursive and takes into account different orders resulting into hierarchies of surrounding contexts.

images (x_i, x_j), provides us with a similarity measure. Our goal is to design $k(x_i, x_j)$ by taking into account the intrinsic properties of x_i, x_j and also their social links, i.e., the set of images which are linked 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 $\mathcal{V} = \{v_1, \dots, v_n\}$ correspond to pairs $\{(x_i, \psi_f(x_i))\}_i$ and edges $\mathcal{E} = \{e_{i,j,\omega}\}$ are the set of labeled connections of $\mathcal{G}$. In the above definition, $\psi_f(x_i)$ corresponds to the descriptor of x_i 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 label for instance two images are linked when they share the same semantic (actually keywords), owners, GPS locations, etc.
Introduce

$$\mathcal{N}^{\omega}(x_i) = \big\{x_j : (x_i, x_j, \omega) \in \mathcal{E}\big\}$$

Notice that the definition of the neighborhood in this paper is different from the one proposed in Sahbi et al. (2008a), as the latter provides us only with a set of neighbors $\mathcal{N}(x)$ around x which are not segmented into different parts. In this work $\mathcal{N}(x) = \cup_{\omega \in \mathcal{L}} \mathcal{N}^{w}(x)$, so our new definition of neighborhoods $\{\mathcal{N}^{w}(x)\}_w$ reflects the co-occurrence of different images with particular words or connection types (see Fig 1, B).

2.2 Context Based Similarity 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 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 $\mathbf{D}_{x,x'} = d(x, x')$, $(d(x, x') = \|\psi_f(x) - \psi_f(x')\|_2)$. We propose to use the kernel on $\mathcal{X}$ defined by solving

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

Here the operations log and $\leq$ are applied individually to every entry of the matrix (for instance, $\log \mathbf{K}$ is the matrix with $(\log \mathbf{K})_{x,x'} = \log k(x, x')$), $\|\cdot\|_1$ is the "entrywise" L_1-norm (i.e., the sum of the absolute values of the matrix coefficients) and Tr denotes matrix trace. The first term, in the above constrained minimization problem, measures the quality of matching two descriptors $\psi_f(x)$, $\psi_f(x')$. In the case of visual features, this is considered as the distance, $d(x, x')$, between the visual descriptors (color, texture, 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 aligned samples, the probability distribution $\{k(x, x')\}$ should be flat

so the negative of the entropy is minimized. This term also helps defining a simple solution and solving the constrained minimization problem easily. The third term is a neighborhood criterion which considers that a high value of $k(x, x')$ should imply high kernel values in the neighborhoods $\mathcal{N}^{\omega}(x)$ and $\mathcal{N}^{\omega}(x')$. This criterion makes it possible to consider the context and social links of each sample in the matching process.
We formulate the minimization problem by adding an equality constraint and bounds which ensure a normalization of the kernel values and allow to see $\{k(x, x')/\sum_{x,x' \in \mathcal{X}} k(x, x')\}$ as a joint probability distribution on $\mathcal{X} \times \mathcal{X}$ (or P-Kernel (Haussler, 1999)).

2.3 Solution

Proposition 1 *Let* $\mathbf{u}$ *denote the matrix of ones and introduce*

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

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

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

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

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

$$\mathbf{K}^{(t)} = \frac{G(\mathbf{K}^{(t-1)})}{\|G(\mathbf{K}^{(t-1)})\|_1}, \tag{4}$$

with

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

and

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

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

$$\|\mathbf{K}^{(t)} - \tilde{\mathbf{K}}\|_1 \leq L^t \|\mathbf{K}^{(0)} - \tilde{\mathbf{K}}\|_1. \tag{6}$$

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

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

To have partitioned the neighborhood into several (typed) links corresponding to different degrees of proximity (as shown in Fig. 1, B) has lead to significant improvements.

On a theoretical viewpoint, it allows us to consider a larger α (since the constraint (2) becomes easier to satisfy with partitioned neighborhood), and apparently a more positively influencing context-dependent term (last term in (1)).
Proof see appendix. ■

We will now show the unicity of the solution of this fixed point equation (4)

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

As a consequence of this lemma, as soon as we have $L = \zeta \exp(\zeta) < 1$, the fixed point equation (4) admits a unique solution $\tilde{\mathbf{K}}$, and Inequality (6) holds.
Proof see appendix. ■

2.4 Reproducing Kernel Hilbert Space

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

Proposition 3 *The context-dependent kernels on $\mathcal{X}$ defined in Proposition (1) by the matrices $\tilde{\mathbf{K}}$ and $\mathbf{K}^{(t)}$, $t \geq 0$, are positive definite.*

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

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

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

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

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

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

3. Benchmarking

3.1 Databases and Settings

In order to show the extra advantage of our kernel with respect to the use of "add-hoc" context-free similarities, we evaluated SNK on classic databases (Corel) as well as social network data from Flickr. Both sets are challenging; the first one, relatively small, contains 6683 images belonging to 200 categories (see Fig 3, top) while the second one is larger and contains 24.999 images downloaded from Flickr[3] and belong to 9 classes (see Fig 3, bottom). Images of these sets contain on average 7 labels. In both sets, social hyper-links are defined between two images if and only if they share common labels. Notice that only labels appearing less than M times are used in order to define links between images so this will reduce the effect of commonly used and useless labels.
Without extensive tuning, we found that any small value of M (in practice $M = 10$) substantially improves the retrieval performance of SNK w.r.t to classic similarities (see Section 3.2). Large values of M were not tested as this results into lots of interconnections in $\mathcal{G}$ and makes the approach computationally intractable.

Each image in Corel and Flickr is processed in order to extract a battery of visual descriptors including HSV, Laplacian RGB, Edge Orientation Histograms (EOH) and Hough coefficients, all concatenated together in order to form a large feature vector. Afterwards, principal component analysis (PCA) is applied on the whole sets (Corel and Flickr) in order to reduce dimensionality by taking 98% of the statistical variance resulting into a feature space of 99 dimensions for Corel and 104 for Flickr.

As a matter of comparison, text features are also combined with visual ones. First, images characterized by their bags of keywords, are mapped using the TF/IDF (term frequency-inverse document frequency) feature vectors then concatenated with the underlying visual features. After the application of PCA on the whole Corel (resp. Flickr) combined features, only 43 (resp. 44) coefficients are kept in order to preserve 98% of the statistical variance of the data. In the remainder of this paper, performances will be reported using a variant of precision/recall, denoted $\mathbf{PR}(i),\ i = 1, ..., 10$, and defined as the expectation of the fraction of relevant images among the top i, here the expectation is with respect to all the possible image queries in the social network.

[3] These images were downloaded and annotated by the ImageClef 2009 challenge organizers.

3.2 Ranking

For both Corel and Flickr, we first build the SNK Gram matrices $\mathbf{K}^{(t)},\ t = 0, 1, \dots$ containing all the cross-similarities between images, then retrieval is achieved by submitting all the possible images as queries. Results are ranked according to relevance as an increasing function of SNK so images ranked among the top are more likely to belong to the same class as the query.
We evaluated our kernel using a power assist Gaussian initialization i.e., $\mathbf{K}^{(0)}_{x,x'} = \exp(-\|\psi_f(x) - \psi_f(x')\|^2/\beta)$. Our goal is to show the improvement brought when using $\mathbf{K}^{(t)}$, $t \in \mathbb{N}^+$, so we tested it against the standard context-free similarity kernel $\mathbf{K}^{(t)},\ t = 0$.
The setting of β is performed by maximizing the performance of the Gaussian kernel as the latter corresponds to the left-hand side (and baseline form) of $\mathbf{K}^{(t)}$, i.e., when $\alpha = 0$. For both Corel and Flickr, we found that the best performances are achieved for $\beta = 10^3$ (see Table 1) and this also guarantees condition (3). The influence (and the performance) of the right-hand side of $\mathbf{K}^{(t)},\ \alpha \neq 0$ increases as α increases (see Table 2), nevertheless and as shown earlier, the convergence of $\mathbf{K}^{(t)}$ to a fixed point is guaranteed only if (2) is satisfied. Therefore it is obvious that α should be set to the highest possible value which also satisfies condition (2).

Diagrams in figure (2) show the **PR** performance on both Corel and Flickr for different iterations; we clearly see the out-performance and the improvement of SNK (i.e., $\mathbf{K}^{(t)},\ t \in \mathbb{N}^+$) with respect to the context-free kernel $\mathbf{K}^{(0)}$. In all cases the improvement brought by our kernel is clear and consistent. Due to the hardness of retrieval tasks on these generic image datasets, the absolute performance reported on these tables are of course smaller than those known for specific databases (such as faces or handwritten characters), nevertheless our main point is to show that SNK consistently improves the results on these generic sets. We have also shown baseline (context-free) results using visual and also combined text/visual features in order to corroborate 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 (see Tables 1,2 and diagrams 2, mainly comparison between visual (Vis) and combined text/visual (Tex/Vis) features).

4. Discussion

Relation to diffusion map. One can easily show that SNK may also capture the structure and the topology of social networks through diffusion map (Lafon et al., 2006). Considering $\mathbf{P}_\omega = \mathbf{P}_{\omega'} = \mathbf{P}$, $\forall \omega, \omega' \in \mathcal{L}$, $|\mathcal{L}| = m$, $\beta \gg 2\alpha m$[4] and using the first order Taylor expansion one may approximate the kernel $\mathbf{K}^{(t)}$ by

$$-\sum_{k=0}^{t-1} \frac{1}{\beta}\left(\frac{2m\alpha}{\beta}\right)^k \mathbf{P}^{(k)}\mathbf{D}\mathbf{P}^{(k)\prime} + \left(\frac{2m\alpha}{\beta}\right)^t \mathbf{P}^{(t)}\mathbf{K}^{(0)}\mathbf{P}^{(t)\prime} \tag{7}$$

[4] This assumption is also well supported by the good performance of our kernel when β takes high values as shown in tables (1, 2).

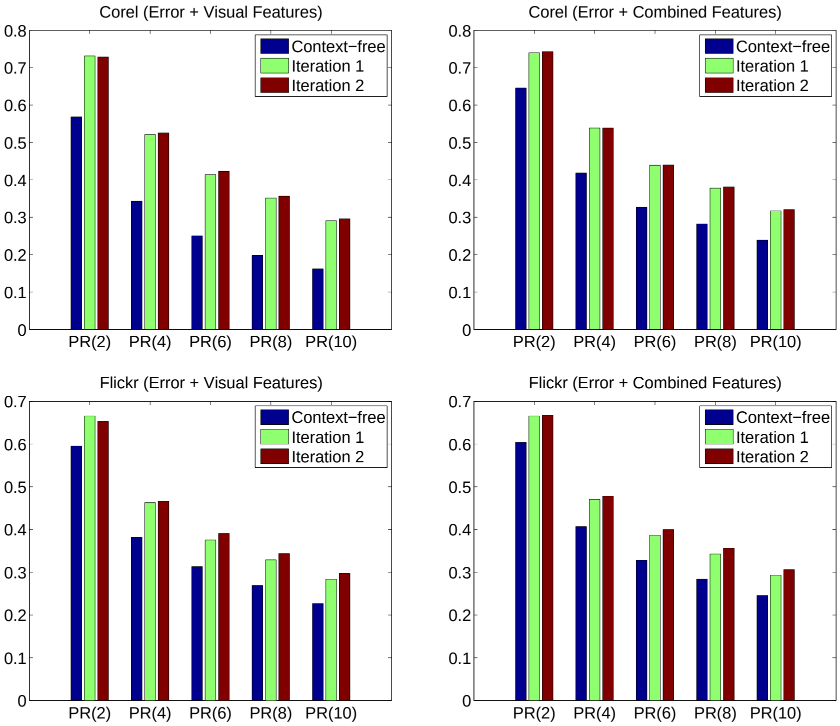

Figure 2: This figure shows different PR results (precision for different recalls) on both Corel and Flickr using Visual and Combined Text/Visual Features. In all these cases, the improvement is always consistent.

Table 1: This table shows the evolution of the PR(10) measure w.r.t. the parameter β. Different feature vectors are used for comparisons including visual (Vis) and combined text/visual (Com).

	Corel		**Flickr**	
(log β)	**Vis**	**Com Tex/Vis**	**Vis**	**Com Tex/Vis**
-2	10.2	10.2	10.0	10.0
-1	10.2	10.2	10.0	10.0
0	12.5	16.5	14.5	19.1
+1	16.0	23.3	22.3	24.4
+2	16.2	23.7	22.6	24.5
+3	**16.2**	**23.8**	**22.6**	**24.5**
PR(10) in (%)				

Table 2: This table shows the evolution of PR(10) measure w.r.t. the parameter α after convergence ($\log \beta = +3$). Different feature vectors are used for comparisons including visual (Vis) and combined text/visual (Com).

	Corel		**Flickr**	
(log α/β)	**Vis**	**Com Tex/Vis**	**Vis**	**Com Tex/Vis**
-3	16.2	23.8	22.6	24.5
-2	16.9	24.6	22.6	24.7
-1	23.8	29.7	25.6	28.1
0	**29.6**	**32.0**	**29.7**	**30.5**
+1	NC	NC	NC	NC
PR(10) in (%), NC stands for not convergent.				

Let $\mathbf{P}_{ij} = P_1(j|i)$ denotes the probability of a 1-step walk from from a node v_i to v_j in $\mathcal{G}$. In the context of diffusion map (Lafon et al., 2006), the idea is to represent higher order walks by taking powers of $\mathbf{P}$, i.e., $\mathbf{P}^{(k)} = \mathbf{P}^{(k-1)}\mathbf{P}$. $\mathbf{P}^{(k)}$ is then the k-step random walk graph Laplacian which models a Markovian process where the conditional k-step transition likelihood is the sum of all the possible k-1 steps linking v_i to v_j ($\mathbf{P}^{(k)}_{ij} = P_k(j|i) = \sum_{\ell=1}^{n} P_{k-1}(\ell|i)P_1(j|\ell)$). In this definition, k acts as a scale factor that increases the local influence of the context when designing SNK. For a given t, the right-hand side of (7) is the "t-step" similarity between vectors *embedded into the manifold space related to diffusion map* while the left-hand side takes into account the intrinsic visual similarity which is independent from the structure of the social network. Put differently, at the convergence stage, SNK balances the visual aspects (features) of images and the underlying manifold topology in the social networks.

Performance. In almost all cases, one iteration was sufficient in order to outperform the Gaussian kernel (baseline). Experiments clearly show that the performance of the latter

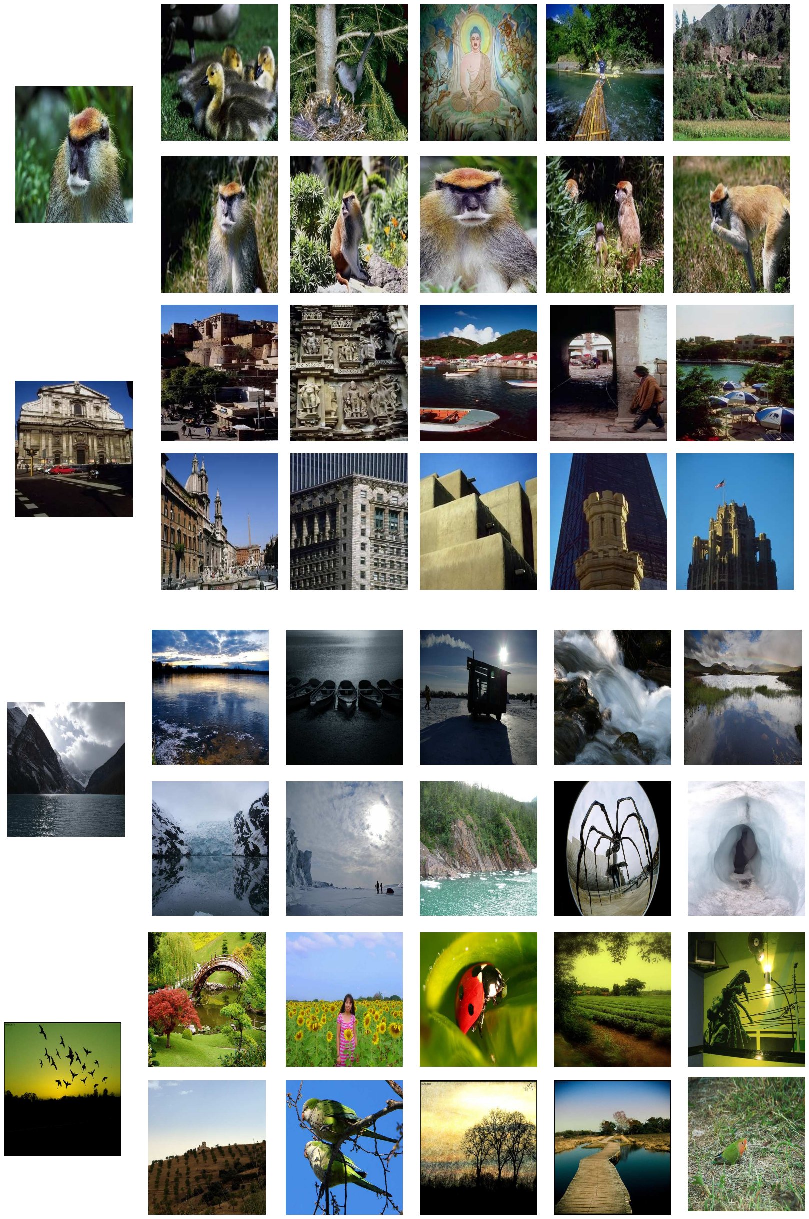

Figure 3: This figure shows ranked results for different image queries in the left-hand side. For each case, upper (resp. lower) images correspond to context-free (resp. dependent) kernel ranking. The two first queries are taken from Corel while the two others from Flickr.

can be consistently improved by including context of images in the SN. Even though results are reported on ranking and retrieval tasks, our kernel might be extended to handle other machine learning problems including classification and regression which are out of the scope of this paper. Furthermore, it is known that good ranking performances of kernels imply good classification/regression results.

Speedup. Finally, one current limitation of SNK (when t takes large values), resides in its computational complexity specially for large scale databases. Assuming $\mathbf{K}^{(t-1)}$ known, for a given pair x, x', the worst complexity is $O(\max(mN^2, ms))$ where s is the dimension of the feature space, $m = |\mathcal{L}|$ (label vocabulary size) and $N = \max_{x,x',\omega}\{|\mathcal{N}^{\omega}(x)|, |\mathcal{N}^{\omega}(x')|\}$. It is clear enough that when $N < \sqrt{s}$, the complexity of evaluating our kernel is equivalent to usual context-free ones such as the Gaussian. That's why we eliminated highly frequent tags when building $\mathcal{G}$; not only because they might not be informative but also in order to decrease interconnections in $\mathcal{G}$ (and also N) and the overhead due to the right-hand side term of SNK. This speedup is still not sufficient mainly when querying very large scale SNs. As a future work, and taking benefit from the Mercer condition, we are currently extending our SNK similarity in order to handle very large scale databases using lossless acceleration techniques.

5. Conclusion

We introduced in this work a novel approach for kernel design dedicated to social networks. The strength of this method resides *both* in the inclusion of social links which provide valuable informations about image relationships and also in the way context is included in kernel design thereby improving ranking and retrieval performances consistently.

Extensions of this work include the use of ontologies in order to enrich social links by looking for synonymous or related concepts. Other future work will tackle acceleration techniques which help reducing processing time and handling very large scale social networks.

Appendix

Proof [Proof of Proposition 1] Introduce the function

$$\begin{aligned} F : \mathbf{K} \mapsto \operatorname{Tr}\big(\mathbf{K}\ \mathbf{D}'\big) \ + \ \beta \operatorname{Tr}\big(\mathbf{K}\ \log\ \mathbf{K}'\big) \\ - \alpha \sum_{\omega} \operatorname{Tr}\big(\mathbf{K}\ \mathbf{P}_{\omega}\ \mathbf{K}'\ \mathbf{P}'_{\omega}\big). \end{aligned}$$

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

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

hence when

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

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

$$\begin{aligned} &\mathbf{K} = \frac{G(\mathbf{K})}{\|G(\mathbf{K})\|_1}, \\ &\text{with} \\ &G(\mathbf{K}) = \exp\Big\{ -\frac{\mathbf{D}}{\beta} + \frac{\alpha}{\beta} \sum_{\omega} \big(\mathbf{P}_\omega \mathbf{K} \mathbf{P}'_\omega + \mathbf{P}'_\omega \mathbf{K} \mathbf{P}_\omega\big) \Big\}, \end{aligned} \tag{8}$$

where the function exp is applied individually to every entry of the matrix. ■

Proof [Proof of Lemma 2] Let $\mathbf{K}_1$ and $\mathbf{K}_2$ be two matrices in $\mathcal{B}$. Introduce $\mathbf{G}_1 = G(\mathbf{K}_1)$ and $\mathbf{G}_2 = G(\mathbf{K}_2)$. We have

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

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

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

Therefore we have

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

which implies

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

Now we trivially have

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

hence we obtain

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

Plugging this inequality into (9), we get

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

■

References

M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. *In the proceedings of CHI*, 2007.

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.

Thomas Arni, Paul Clough, Mark Sanderson, and Michael Grubinger. Overview of the imageclefphoto 2008 photographic retrieval task. *Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark*, 2008.

F. Bach. Graph kernels between point clouds. *In proceeding of the International Conference on Machine Learning*, 2008.

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

J. Bian, Y. Liu, E. Agichtein, and H. Zha. A few bad votes too many? towards robust ranking in social media. *in the WWW 2008 workshop on Adversarial Information Retrieval (AIRWeb)*, 2008.

ECIR. Ecir workshop on social networks, http://jmgomezhidalgo.blogspot.com/2009/01/ecir-workshop-on-information-retrieval.html. *In the European Conference on Social Networks*, 2009.

M. Franke, B. Hoser, and J. Schröder. Enlarging personal networks through transitive clustering. *In Sunbelt XXVII International Social Network Conference*, 2007.

C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. *In the proceedings of IEEE conference on Computer vision and Pattern Recognition CVPR*, 2008.

S. Geman and M. Johnson. Dynamic programming for parsing and estimation of stochastic unification-based grammars. *In the proceedings of ACL*, 2002.

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

S. Hoi, W. Liu, and S.F. Chang. Semi-supervised distance metric learning for collaborative image retrieval. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) Anchorage, Alaska, USA*, 2008.

B. Hoser. Information retrieval vs knowledge retrieval: A sn perspective. *In CIS Web*, 2009.

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

Y. Jin and S. Geman. Context and hierarchy in a probabilistic image model. *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2:2145–2152, 2006.

L. Kirchhoff, K. Stanoevska-Slabeva, T. Nicolai, and M. Fleck. Using social network analysis to enhance information retrieval systems. *http://www.alexandria.unisg.ch/Publications/46444*, 2008.

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

Stephane Lafon, Yosi Keller, and Ronald R. Coifman. Data fusion and multi-cue data matching by diffusion maps. *IEEE Trans. Pattern Anal. Mach. Intell*, 28(11):1784–1797, 2006.

X. Li, C. Snoek, and M. Worring. Learning tag relevance by neighbor voting for social image retrieval. *MIR*, 2008.

D. Ritendra, D. Joshi, J. Li, and J.Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 2008.

S. Roweis and L. Saul. Non linear dimensionality reduction by locally linear embedding. *Science*, 290(5500), 2000.

B. Russell, A. Torralba, K. Murphy, and W. T. Freeman. Label-me: a database and web-based tool for image annotation. *IJCV*, 2008.

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

H. Sahbi, P. Etyngier, J.Y. Audibert, and R. Keriven. Manifold learning using robust graph laplacian for interactive image search. *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008b.

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

A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

Za. Stone, T. Zickler, and T. Darrell. Auto-tagging facebook: Social network context improves photo annotation. *in IVW*, 2008.

J. Wang, T. Jebara, and S.F. Chang. Graph transduction via alternating minimization. *International Conference on Machine Learning (ICML), Helsinki*, 2008.

WSIRTEL. 1st workshop on social information retrieval for technology-enhanced learning, http://prolearnsummerschool.wordpress.com/category/social-information-retrieval/. *In the 2nd European Conference on Technology Enhanced Learning (EC-TEL07), Crete, Greece,September 17-20*, 2007.

D Zhou, J. Bian, S. Zheng, H. Zha, and C.L. Giles. Exploring social annotations for information retrieval. *in the international WWW conference, Beijing, China*, 2008.

S. Zhu and D. Mumford. A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, 2, 2004.

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

TELECOM ParisTech

Institut TELECOM - membre de ParisTech

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

Département TSI