Context Dependent SVMs for Interconnected Image Network Annotation

Hichem SAHBI
CNRS TELECOM ParisTech,
Paris, France
hichem.sahbi@telecom-paristech.fr

Xi LI
CNRS TELECOM ParisTech, Paris, France
& NLPR CASIA, Beijing, China
lixichinanlpr@gmail.com

ABSTRACT

The exponential growth of interconnected networks, such as Flickr, currently makes them the standard way to share and explore data where users put contents and refer to others. These interconnections create valuable information in order to enhance the performance of many tasks in information retrieval including ranking and annotation. We introduce in this paper a novel image annotation framework 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 a variational framework which helps designing this function using both intrinsic features and the underlying contextual information. This function also converges 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.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2 [**Artificial Intelligence**]: Learning—*Concept Learning*

General Terms

Algorithms, Theory, Experimentation

Keywords

Kernel design, support vector machines, context, image retrieval, interconnected and social networks

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 Content-Based Image Annotation and Search; where images are represented using low-level visual features (color, texture, shape, etc.), and annotated by analyzing those features in order to extract keywords. Conventionally, image annotation is converted into a classification problem. Existing state of the art methods (for instance [2, 12]) treat each keyword or concept 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 [12], latent Dirichlet allocation [1], probabilistic latent semantic analysis [16], and support vector machines [7]. The aforementioned annotation methods may also be categorized into two branches; region-based requiring a preliminary step of image segmentation [12, 13], and holistic [8, 21] 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. Due to the semantic gap, they are unable to fully explore the semantic information inside images. Another class of annotation methods has then 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) [8, 10] and its variants [5], concept propagation in [14], graph based semi-supervised inference [13] and others inspired from machine translation [4]. Other existing annotation methods focus on how to define an effective distance measure for exploring the semantic relationships between concepts in large scale databases; including the Normalized Google similarity Distance, in [3], which measures the semantic correlations derived from counts returned by Google's search engine for a given set of keywords. Following the idea of [3], the Flickr distance [23] 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 WordNet [18, 15]) in order to enrich annotations [22]. In [20, 9], the semantic ontology information is integrated in the post processing stage in order to further refine initial annotations.

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 [19]. Support vector machines basically require the design of similarity measures, also referred to as *kernels*, which should provide

Copyright 2010 Association for Computing Machinery. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of the French Government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.
MM'10, October 25–29, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-60558-933-6/10/10 ...$10.00.

high values when two images share similar 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 [19], 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 framework 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 [6, 11, 24]) 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 [19]. This type of kernels will be referred to as context-dependent (CDK) while those relying only on the intrinsic visual 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. 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.

2. KERNEL DESIGN

Let us consider $\mathcal{X} = \{x_1, \ldots, x_n\}$ as a finite set of images drawn from an existing (but unknown) probability distribution. 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) $\mathcal{N}^\omega(x_i) = \{x_j : (x_i, x_j, \omega) \in \mathcal{E}\}$. This definition of context $\{\mathcal{N}^w(x)\}_w$ reflects the co-occurrence of different images with particular connection types (again defined using tags).

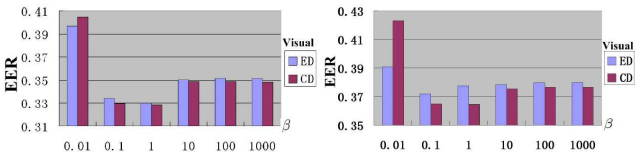

Figure 1: This figure shows the annotation performances of different βs under different metrics on the MIRFLICKR-25000 (left) and NUSWIDE datasets (right).

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

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

Here $\alpha,\ \beta \geq 0$ and the operations log (neperian) 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 "entry-wise" L_1-norm (i.e., the sum of the absolute values of the matrix coefficients) and $\operatorname{Tr}(\cdot)$ denotes matrix trace. 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 on $\mathcal{X} \times \mathcal{X}$ (or P-Kernel).

2.3 Solution

The above optimization problem admits a solution $\tilde{\mathbf{K}}$, which is the limit of the context-dependent kernels $\mathbf{K}^{(t)} = \frac{G(\mathbf{K}^{(t-1)})}{\|G(\mathbf{K}^{(t-1)})\|_1}$, with $G(\mathbf{K}) = \exp\big\{-\frac{\mathbf{D}}{\beta} + \frac{\alpha}{\beta}\sum_\omega \big(\mathbf{P}_\omega \mathbf{K} \mathbf{P}'_\omega + \mathbf{P}'_\omega \mathbf{K} \mathbf{P}_\omega\big)\big\}$, and $\mathbf{K}^{(0)} = \frac{\exp(-\mathbf{D}/\beta)}{\|\exp(-\mathbf{D}/\beta)\|_1}$. By taking small enough α, convergence of this kernel to a fixed point is satisfied (see [17]). 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 positive definite fixed point may be found in [17]. Note that positive definiteness of a kernel k guarantees that the underlying Gram matrix $\mathbf{K}$ is positive (semi-)definite, i.e., the existence of a mapping function in some Reproducing Kernel Hilbert Space (RKHS), such that k is written as a dot product in the RKHS (see for instance [19]).

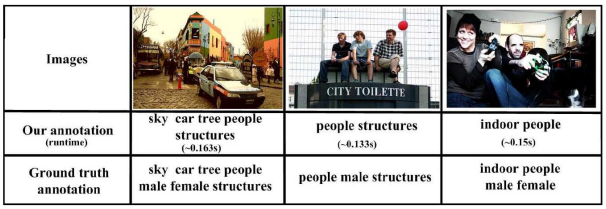

Figure 2: This table shows comparisons of ground truth and the best context-dependent kernel (HI+CDK) annotations on three images from MIRFLICKR-25000.

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

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.

3.2 Hold-Out Generalization and Comparison

We use six power assist settings as $\mathbf{K}^{(0)}$; Linear: $\langle \psi_f(x), \psi_f(x') \rangle$, Polynomial: $\langle \psi_f(x), \psi_f(x') \rangle^2$, RBF: $\exp(-\frac{1}{\beta}\|\psi_f(x) - \psi_f(x')\|_2)$, Histogram intersection: $\sum_i \min(\psi_f(x)_i, \psi_f(x')_i)$, Chisquare: $1 - \frac{1}{2}\sum_i \frac{(\psi_f(x)_i - \psi_f(x')_i)^2}{(\psi_f(x)_i + \psi_f(x')_i)}$, Sigmoid: $\tanh(\langle \psi_f(x), \psi_f(x') \rangle)$, and the underlying CDK kernels $\mathbf{K}^{(t)}$, $t \in \mathbb{N}^+$ 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[3] for each concept in 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). The latter is the balanced generalization error which equally weights the positive and the negative errors. The smaller the EER, the better the annotation performance. This measure is 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) $d(x, x') = \|\psi_f(x) - \psi_f(x')\|_2$; and (ii) Chisquare distance (CD) $d(x, x') = \frac{1}{2}\sum_i \frac{(\psi_f(x)_i - \psi_f(x')_i)^2}{(\psi_f(x)_i + \psi_f(x')_i)}$. Based on the above two metrics, we tuned the scale parameter β to achieve the best EER annotation performances. Fig. 1 shows the EER annotation results of different βs under different metrics on the MIRFLICKR-25000 and the NUSWIDE datasets. Clearly, we found that the best performances are achieved on both MIRFLICKR-25000 and NUSWIDE datasets when $\beta = 1$ using the CD metric.

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. 3), nevertheless and as shown in [17], 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 [17]). 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 $2^{1}\beta$.

Comparison. Fig. 3 shows the annotation performance (EER) 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 for MIRFLICKR-25000 and 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). According to Fig. 3, performances of context-dependent kernels are mostly better than those of context-free ones, with just two iterations (i.e., $t \geq 2$). Standard deviations are also shown with respect to different threshold intervals $[T_D, T_U]$. Further illustrations (see Fig. 2), 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. It is clear that our proposed method achieves reasonable annotation results.

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

[1] http://press.liacs.nl/mirflickr/

[2] http://lms.comp.nus.edu.sg/research/NUSWIDE.htm

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

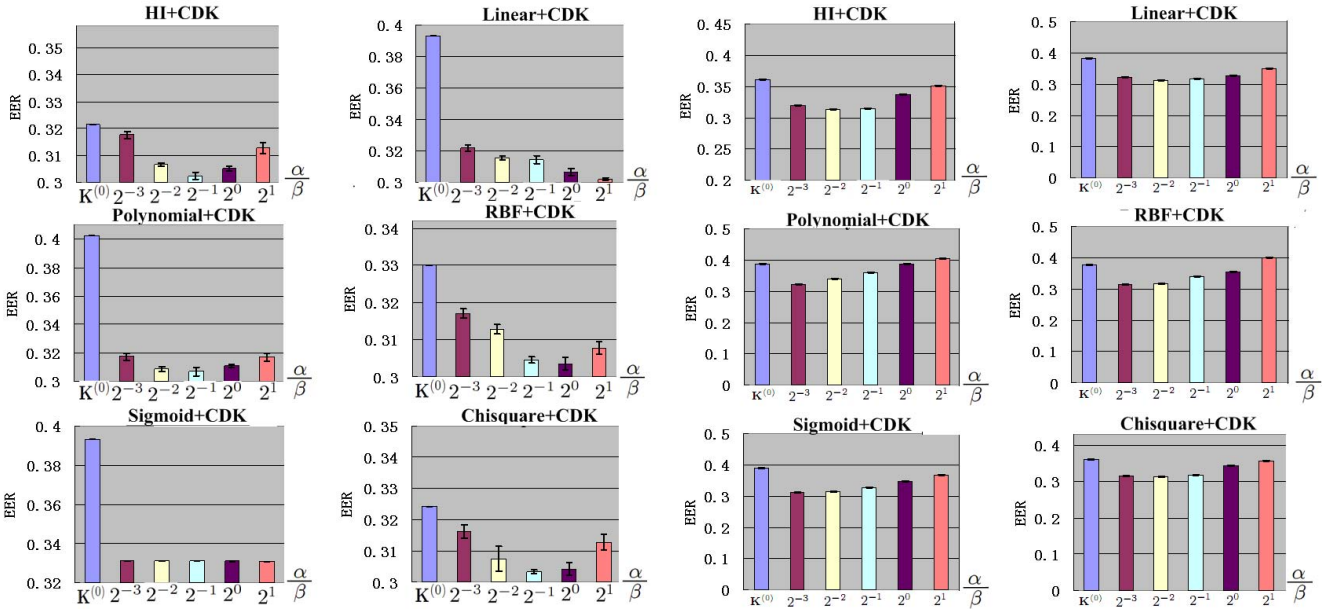

Figure 3: This figure shows the performances of annotation on the MIRFLICKR-25000 dataset (left) and NUSWIDE (right) (with $\beta = 1$). It includes EER annotation performances of six context-dependent kernels based on visual features. Compared with the underlying baseline kernels, the six (best) context-dependent kernels achieve a relative gain of respectively 5.91%, 23.09%, 23.69%, 8.03%, 17.34%, and 6.47% for MIRFLICKR-25000 and 12.98%, 18.28%, 12.76%, 16.18%, 19.23%, and 12.94% for NUSWIDE.

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.

5. ACKNOWLEDGMENTS

This work was supported by the French ANR project AVEIR.

6. REFERENCES

[1] K. Barnard, P.Duygululu, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *JMLR*, 2003.

[2] G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *CVPR*, 2005.

[3] R. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 2007.

[4] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. of ECCV*, 2002.

[5] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Proc. of ICCV*, pages 1002–1009, 2004.

[6] A. Gallagher, C. Neustaedter, L. Cao, J. Luo, and T. Chen. Image annotation using personal calendars as context. In *ACM MM*, 2008.

[7] Y. Gao, J. Fan, X. Xue, and R. Jain. Automatic image annotation by incorporating feature hierarchy and boosting to scale up svm classifiers. In *Proc. of ACM MM*, 2006.

[8] J. Jeon, V. Lavrenko, and R.Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proc. of ACM SIGIR*, pages 119–126, 2003.

[9] Y. Jin, L. Khan, L. Wang, and M. Awad. Image annotations by combining multiple evidence & wordnet. In *Proc. of ACM Multimedia*, pages 706–715, 2005.

[10] J.Liu, B.Wang, M.Li, Z.Li, W.Ma, H.Lu, and S.Ma. Dual cross-media relevance model for image annotation. In *Proc. of ACM MM*, pages 605–614, 2007.

[11] C. L., L. J., and H. T.S. Annotating photo collection by label propagation according to multiple similarity cues. In *Proc. of ACM Multimedia*, 2008.

[12] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on PAMI.*, 25(9):1075–1088, 2003.

[13] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma. Image annotation via graph learning. *Pattern Recognition*, 42(2):218–228, 2009.

[14] Z. Lu, H. H. Ip, and Q. He. Context-based multi-label image annotation. In *Proc. of ACM CIVR*, 2009.

[15] G. A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[16] F. Monay and D. GaticaPerez. Plsa-based image autoannotation: Constraining the latent space. In *Proc. of ACM MM*, 2004.

[17] H. Sahbi and J.-Y. Audibert. Social network kernels for image ranking and retrieval. In *Technical Report, N 2009D009, TELECOM ParisTech*, March 2009.

[18] M. Srikanth, J. Varner, M. Bowden, and D. Moldovan. Exploiting ontologies for automatic image annotation. In *Proc. of ACM SIGIR*, pages 552–558, 2005.

[19] V. Vapnik. *Statistical Learning Theory*. A Wiley-Interscience Pub, 1998.

[20] C. Wang, F. Jing, L. Zhang, and H. J. Zhang. Image annotation refinement using random walk with restarts. In *Proc. of ACM Multimedia*, pages 647–650, 2006.

[21] C. Wang, S. Yan, L. Zhang, and H. Zhang. Multi-label sparse coding for automatic image annotation. In *Proc. of CVPR*, 2009.

[22] Y. Wang and S. Gong. Translating topics to words for image annotation. In *ACM CIKM*, 2007.

[23] L. Wu, X. Hua, N. Yu, W. Ma, and S. Li. Flickr distance. In *Proc. of ACM Multimedia*, 2008.

[24] Y. Yang, P. Wu, C. Lee, K. Lin, W. Hsu, and H. Chen. Contextseer: Context search and recommendation at query time for shared consumer photos. In *Proc. of ACM Multimedia*, 2008.