# Learning an Image-Word Embedding for Image Auto-Annotation on the Nonlinear Latent Space

Wei Liu
Department of Information Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong
wliu5@ie.cuhk.edu.hk

Xiaoou Tang
Microsoft Research Asia
Beijing, China
xitang@microsoft.com

## ABSTRACT

Latent Semantic Analysis (LSA) has shown encouraging performance for the problem of unsupervised image automatic annotation. LSA conducts annotation by keywords propagation on a linear Latent Space, which accounts for the underlying semantic structure of word and image features. In this paper, we formulate a more general nonlinear model, called Nonlinear Latent Space model, to reveal the latent variables of word and visual features more precisely. Instead of the basic propagation strategy, we present a novel inference strategy for image annotation via Image-Word Embedding (IWE). IWE simultaneously embeds images and words and captures the dependencies between them from a probabilistic viewpoint. Experiments show that IWE-based annotation on the nonlinear latent space outperforms previous unsupervised annotation methods.

## Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*

## General Terms

Algorithms, Theory

## Keywords

auto-annotation of images, semantic indexing, nonlinear latent space, image-word embedding

## 1. INTRODUCTION

The potential value of large image collections fully depends on effective methods for access and search. When adequate annotations are available, searching image collections is intuitive. Moreover, image users often prefer to formulate intuitive text-based queries to retrieve relevant images, which requires the annotation of each image belonging to the collection. Since off-line image annotation is laborious and expensive, automatic image annotation has thus emerged as one of the key research areas in multimedia information retrieval [1][2][3][7]. Image auto-annotation performs automatic association of images with words that describe the image content such as "water" or concepts such as "sunset".

The state-of-the-art techniques for image auto-annotation can be grouped into two categories. The first one looks upon annotation as a supervised learning problem, and associates words to images by first defining classes. Each class corresponds to a word [3], or a set of words define a concept [7]. After training of each visual class model with manually labeled images, each image will be classified into one or more classes, annotation is hence attained by propagating the corresponding class words. The second one addresses image auto-annotation as a unsupervised problem. It attempts to discover the statistical links between visual features and words on an unsupervised basis through estimating the joint distribution of words and regional image features, and elegantly posing annotation as statistical inference in a graphical model [1].

Motivated by the success of latent space models in text analysis, two commonly used text analysis models, named Latent Sematic Analysis (LSA) [4] and Probabilistic LSA (PLSA) [5] have been applied to image annotation in the literature [8][9]. Monay *et al.* [8] show that annotation by LSA+propagation significantly outperforms annotation by PLSA+inference. So in this paper, we only explore the LSA model and develop novel annotation methods based on it. Although generative probabilistic models for PLSA-based auto-annotation have been proposed in [9], they are too complex (too many parameters unknown) to be generalized.

We first present the nonlinear latent space model as an alternative to existing latent space models [4][8] which are linear. Later a soft inference strategy is performed on the learned nonlinear latent space, which is simple but effective. We call the proposed auto-annotation method *Nonlinear Latent Semantic Analysis (NLSA)*. Second, we learn an *Image-Word Embedding (IWE)*, which takes as input inferred results by NLSA, for a large collection including annotated and non-annotated images. IWE can capture the inherent high-level probabilistic relations within and across the textual (words) and visual (images) modalities. IWE incorporated with NLSA constructs our automatic image annotation framework.

## 2. IMAGE REPRESENTATION

Presumably we refer a more general "vocabulary" for a


*MM'05,* November 6–11, 2005, Singapore.

collection of annotated images over separate sets of all observed textual keywords and visual keywords, which are referred as "terms" in [8]. Annotated images are thus summarized as extended documents combining two complementary modalities, i.e. textual and visual modalities which are both represented in a discrete vector-space form.

**Textual feature: word**. The set of words of an annotated image collection defines a keywords vector-space of dimension $K$($K$ words in total), where each component indexes a particular keyword $W$ that occurs in an image. The textual modality of the $d$-th annotated image $I_d$ is hence represented as a vector $\mathbf{w}_d = (w_d^1, \cdots, w_d^i, \cdots, w_d^L)^T$ of size $L$, where each element $w_d^i$ is the frequency (count) of the corresponding word $W_i$ in "document" $I_d$.

**Visual feature: image**. In line with the successful image representation: 6×6×6 RGB histograms [8], we compute 3 RGB color histograms from three fixed regions (i.e. the image center, and the upper and lower parts in the image)[1]. To keep only the dominant colors, a threshold is used to keep only higher values. So the visual modality of $I_d$ is a vector of size $\bar{L} = 3 * 6^3$: $\mathbf{v}_d = (v_d^1, \cdots, v_d^j, \cdots, v_d^{\bar{L}})^T$.

Then, the concatenated vector represented by $\mathbf{x}_d = (\mathbf{w}_d; \mathbf{v}_d)$ is the feature of the annotated image $I_d$ which is the $d$-th multimedia document in the collection. Now we define a term-by-document matrix $\mathrm{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_d, \cdots, \mathbf{x}_N] \in \Re^{M \times N}$ for an annotated image collection, where $N$ is the number of documents (images) and $M$ is the vocabulary size $L + \bar{L}$.

Each annotated image can be thought as an interaction between textual and visual factors, each factor referring to the other. For instance, an image potentially illustrates hundreds of textual words, while its caption specifies the visual content. As for non-annotated images $I_q$, corresponding representations $\mathbf{x}_q$ also are computed in the full vocabulary vector-space, with all elements corresponding to textual keywords set to zero, i.e. $\mathbf{x}_q = (\mathbf{0}; \mathbf{v}_q)$.

# 3. LSA-BASED ANNOTATION

## 3.1 Latent Space

A classic algorithm arising from linear algebra, LSA decomposes the term-by-document matrix $\mathrm{A} = \mathrm{X}^T \in \Re^{N \times M}$ in three matrices by a truncated Singular Value Decomposition (SVD):

$$\mathrm{A} \cong \mathrm{USV}^T, \tag{1}$$

where $\mathrm{U} \in \Re^{N \times K}$, $\mathrm{S} \in \Re^{K \times K}$ and $\mathrm{V} \in \Re^{M \times K}$. The operation performs the optimal least-square projection of the original space onto a space of reduced dimensionality $K$. The subspace representation has been empirically shown to capture to some degree the semantic relationships across terms in a corpus.

Specifically, V is the latent space basis and all annotated and unannotated image features will be projected on it to compute similarities in the learned latent space. It is easy to realize that the subspace V can also be attained via running PCA on $\mathrm{XX}^T$, i.e. V is the principal component of eigenspace of $\mathrm{XX}^T$. Therefore, latent space model is another description of PCA.

[1]In this paper, we use the simple quantized image representation for the visual feature, other representations such as blobs and LBP will be discussed in our future research.

## 3.2 Propagation vs. Soft Inference

In the latent space, latent features of documents $\mathbf{x}$ are extracted by $\mathrm{V}^T\mathbf{x}$ for whether annotated or non-annotated images in the collection. After the cosine similarity between an unannotated image $\mathbf{x}_q$ and the annotated image corpus is measured, top-$Z$ similar annotated images (documents) are ranked as $\mathbf{x}_{\mathcal{N}(j)} (j = 1, \cdots, Z)$. The annotation is then propagated from the words associated with ranked images.

LSA+propagation has been demonstrated to be rather effective in [8], unfortunately, LSA lacks a clear probabilistic interpretation [5]. Hence image users cannot attach probability to each ranked keyword, also cannot know which annotated word is reliable. Furthermore, the propagation strategy cannot provide a dynamic annotation decision using a threshold level that is often given by users.

To overcome this problem, we propose a simple inference strategy for annotation, named *soft inference*. Here we only consider top-$Z$ ranked documents for any unannotated image $\mathbf{x}_q$, and estimate the posterior distribution over keywords as below

$$P(W_i|\mathbf{x}_q) = C \sum_{j=1}^{Z} \cos(\mathrm{V}^T\mathbf{x}_q, \mathrm{V}^T\mathbf{x}_{\mathcal{N}(j)}) * w_{\mathcal{N}(j)}^i, \tag{2}$$

where $C$ is the normalized constant such that $\sum_i P(W_i|\mathbf{x}_q) = 1$, $\cos(*,*)$ denotes a standard cosine measure in terms of any two vectors. Thereby, words will be predicted for the unannotated image document $\mathbf{x}_q$ with a posterior probability higher than a threshold that may be varied for different users.

# 4. NONLINEAR LATENT SPACE MODEL

LSA builds latent space representation in a linear formulation, which assumes equal relevance for the text and visual modalities. The assumption is not always reliable and is somewhat unreasonable in theory, because textual features and visual features are formed from two very different modalities. Hereby, we develop a nonlinear latent space model to complement the linear one.

## 4.1 Definition of Nonlinear Latent Space

To correlate word and image features with different modalities, integrate them into one unified modality is primary. Here we introduce a nonlinear mapping from the document vector-space $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ to an *implicit feature space* $\mathcal{F}$, i.e. $\phi: \ \mathbf{x} \in \Re^M \longrightarrow \phi(\mathbf{x}) \in \Re^f$, where $f > M$ is the dimension of $\mathcal{F}$ and could be infinite. Because the mapping $\phi$ ensures multi-modal co-occurrences uniform, the same way to LSA, a linear subspace can be computed to capture the relationships across textual and visual terms in corpus. Motivated by analysis in Section 3.1, the latent space basis in $\mathcal{F}$ is leading eigenvectors $\mathbf{U}$ of the following covariance matrix

$$\mathbf{C} = N^{-1} \sum_{n=1}^{N} (\phi(\mathbf{x}_n) - \bar{\phi})(\phi(\mathbf{x}_n) - \bar{\phi})^T = \mathbf{\Phi}\mathrm{YY}^T\mathbf{\Phi}^T = \mathbf{\Psi}\mathbf{\Psi}^T, \tag{3}$$

where mean $\bar{\phi} = \sum_n \phi(\mathbf{x}_n)/N$, $\mathbf{\Phi} = [\phi(\mathbf{x}_1), \cdots, \phi(\mathbf{x}_N)]$, a $N \times N$ constant matrix $\mathrm{Y} = N^{-1/2}(\mathrm{I} - \mathbf{1}\mathbf{1}^T/N)$, $\mathbf{\Psi} = \mathbf{\Phi}\mathrm{Y}$.

Computing $\mathbf{U}$ is consistent to Kernel PCA [10], the implicit feature vector $\phi(\mathbf{x}_n)$ don't need to be computed explicitly, instead it is embodied by computing the inner product of any two vectors in $\mathcal{F}$ utilizing a kernel function,

which is guaranteed by the Mercer's theorem and satisfies $k(\mathbf{x}_1, \mathbf{x}_2) = (\phi(\mathbf{x}_1) \cdot \phi(\mathbf{x}_2)) = \phi(\mathbf{x}_1)^T \phi(\mathbf{x}_2)$. So we define a $N \times N$ matrix

$$\tilde{\mathrm{C}} = \mathbf{\Psi}^T \mathbf{\Psi} = \mathrm{Y}^T \mathbf{\Phi}^T \mathbf{\Phi} \mathrm{Y} = \mathrm{Y}^T \mathrm{K} \mathrm{Y}, \tag{4}$$

where $\mathrm{K} = \mathbf{\Phi}^T \mathbf{\Phi}$ is the $N \times N$ kernel matrix with entry $\mathrm{K}_{(i,j)} = (\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)) = k(\mathbf{x}_i, \mathbf{x}_j)$.

The eigensystem for $\mathbf{C}$ can be derived from $\tilde{\mathrm{C}}$. Suppose that the eigenpairs of $\tilde{\mathrm{C}}$ are $\{(\lambda_n, \mathbf{v}_n)\}_{n=1}^N$ sorted in a non-increasing order w.r.t. $\lambda_n$, then the $K(< N)$ leading eigenvectors $\mathbf{U}_K$ for $\mathbf{C}$ is derived as $\mathbf{\Phi} \mathrm{Y} \mathrm{V}_K \Lambda_K^{-1/2}$, where $\mathrm{V}_K = [\mathbf{v}_1, \cdots, \mathbf{v}_K]$ and $\Lambda_K = Diag[\lambda_1, \cdots, \lambda_K]$. For succinct formulation, assume the $N \times K$ matrix $\mathrm{D} = \mathrm{Y} \mathrm{V}_K \Lambda_K^{-1/2}$ then $\mathbf{U}_K = \mathbf{\Phi} \mathrm{D}$.

By projecting $\phi(\mathbf{x})$ on the basis $\mathbf{U}_K$ latent in $\mathcal{F}$ for any vector $\mathbf{x} \in \Re^M$, we obtain another nonlinear mapping as a explicit form

$$\digamma(\mathbf{x}) = \mathbf{U}_K^T \phi(\mathbf{x}) = \mathrm{D}^T \mathbf{\Phi}^T \phi(\mathbf{x}) = \mathrm{D}^T (k(\mathbf{x}_1, \mathbf{x}), \cdots, k(\mathbf{x}_N, \mathbf{x}))^T. \tag{5}$$

$\digamma$ maps any vector in $\Re^M$ to a $K$-dimensional vector. We denote the mapped feature space $\Omega = \{\digamma(\mathbf{x})\}$ as *Nonlinear Latent Space* which is suitable for multi-modal features, e.g. above concatenated features from textual and visual modalities.

## 4.2 NLSA-based Annotation

In line with LSA and PLSA, we name the text analysis method with nonlinear latent space model as *Nonlinear LSA (NLSA)*. Using annotated images $\mathbf{x}_1, \cdots, \mathbf{x}_N$ as training samples, NLSA learns the nonlinear mapping $\digamma$ for any document $\mathbf{x}$ (including annotated and non-annotated images) as shown in (5).

NLSA-based annotation method consists of two steps: i) similarity calculation in the nonlinear latent space under the nonlinear mapping between the image to be annotated and each annotated image in the corpus, using a standard cosine measure, and ii) soft inference based on top-$Z$ annotated images $\mathbf{x}_{\mathcal{N}(j)} (j = 1, \cdots, Z)$ depending on the similarity rank. There is a discrepancy between current posterior distribution w.r.t. keywords and the previous one in (2), we reformulate it in the nonlinear latent space

$$P(W_i|\mathbf{x}_q) = C \sum_{j=1}^{Z} \cos(\digamma(\mathbf{x}_q), \digamma(\mathbf{x}_{\mathcal{N}(j)})) * w_{\mathcal{N}(j)}^i. \tag{6}$$

# 5. IWE-BASED INFERENCE

To refine annotation by soft inference with (2) and (6), we try to learn and infer the inherent high-level probabilistic relations within and across the textual and visual modalities in a specific embedded space, which will make annotation by inference more reliable.

## 5.1 Image-Word Embedding (IWE)

Motivated by Parametric Embedding [6], it is not necessary to explicitly learn an embedding from $\mathbf{x}$ or $\digamma(\mathbf{x})$. Given a set of class posterior, PE tries to preserve the posterior structure in an embedding space. Here we extend PE to inference-based image annotation, and our inference method is called *Image-Word Embedding (IWE)* which simultaneously embeds both annotated images and their associated words in a low-dimensional space. Impressively, a 2D embedding of words and image features is capable of revealing the high-level probabilistic dependencies within and across textual and visual modalities.

IWE takes as input the following posteriors and priors, which are estimated from the count of each word occurring in each annotated image and the whole corpus respectively

$$P(W_i|\mathbf{x}_n) = w_n^i / \sum_{l=1}^{L} w_n^l, P(W_i) = \sum_{n=1}^{N} w_n^i / \sum_{l=1}^{L} \sum_{n=1}^{N} w_n^l. \tag{7}$$

Then IWE tries to embed annotated images $\mathbf{x}_n$ with coordinates $\mathbf{r}_n$ and words $W_i$ (classes) with mean $\phi_i$, such that $P(W_i|\mathbf{r}_n)$ are approximated as closely as possible by the posterior probabilities from a unit-variance spherical Gaussian mixture model in the embedding space

$$P(W_i|\mathbf{r}_n) = \frac{P(W_i) \exp\{-\frac{1}{2}\|\mathbf{r}_n - \phi_i\|^2\}}{\sum_{l=1}^{L} P(W_l) \exp\{-\frac{1}{2}\|\mathbf{r}_n - \phi_l\|^2\}}, \tag{8}$$

where the embedding-space word conditional distribution $p(\mathbf{r}_n|W_i) = \exp\{-\|\mathbf{r}_n - \phi_i\|^2/2\}$ is from a single spherical Gaussian model.

The degree of correspondence between input probabilities and embedding-space probabilities is measured by sum of Kullback-Leibler (KL) divergences for each annotated images: $\sum_n \mathrm{KL}(P(W_i|\mathbf{x}_n)||P(W_i|\mathbf{r}_n))$. Minimizing this sum w.r.t. $\{P(W_i|\mathbf{r}_n)\}$ is equivalent to minimizing the objective function

$$E(\{\mathbf{r}_n\}, \{\phi_i\}) = -\sum_{i=1}^{L} \sum_{n=1}^{N} P(W_i|\mathbf{x}_n) \log P(W_i|\mathbf{r}_n). \tag{9}$$

Such optimization problem can be solved by employing coordinate descent method, which minimizes $E$ iteratively w.r.t. to $\{\phi_i\}$ or $\{\mathbf{r}_n\}$ while fixing the other set of parameters until convergence. Particularly, the Hessian of $E$ w.r.t. $\{\mathbf{r}_n\}$ is a semi-definite matrix and the globally optimal solution for $\{\mathbf{r}_n\}$ given $\{\phi_i\}$ can be found.

## 5.2 Annotation with IWE

In the testing stage, for any image $\mathbf{x}_q$ to be annotated, we need to minimize the simplified object function w.r.t the embedded coordinate $\mathbf{r}_q$

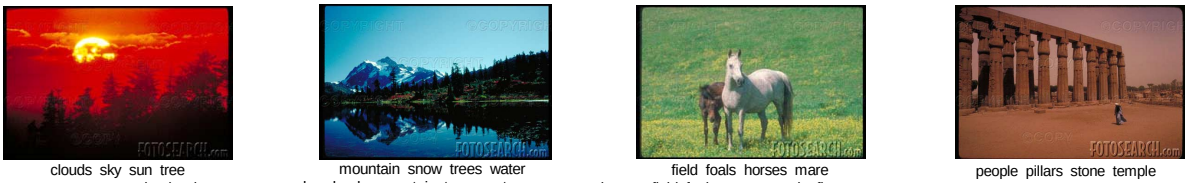
$$E(\mathbf{r}_q) = -\sum_{i=1}^{L} P(W_i|\mathbf{x}_q) \log P(W_i|\mathbf{r}_q). \tag{10}$$

With $\{P(W_i|\mathbf{x}_q)\}$ learned by soft inference (6) and $\{\phi_i\}$ learned in the above training stage, the derivative vector of $E(\mathbf{r}_q)$ is $dE/d\mathbf{r}_q = \sum_{i=1}^{L} (P(W_i|\mathbf{x}_q) - P(W_i|\mathbf{r}_q))(\mathbf{r}_q - \phi_i)$. The optimization is very fast to converge especially for a small dimension, e.g. 2, of the embedding space.

Once the solution $\mathbf{r}_q$ is used to compute the refined posterior distribution $P(W_i|\mathbf{r}_q) (i = 1, \cdots, L)$ in the embedding space instead of $P(W_i|\mathbf{x}_q)$, we create a robust annotation over the full keyword vocabulary by varying a threshold level and predicting the words with a posterior probability higher than this selected threshold.

# 6. EXPERIMENTS

Since most recent image annotation works are performed on the well known image database, the Corel image collection, we use images from it as experimental data as well. Specifically, 10 different subsets are sampled from 80 Corel



clouds sky sun tree
water sun tree sunset clouds sky grass
sun sunset clouds tree sky water
sun sunset clouds sky tree

mountain snow trees water
sky clouds mountain trees water grass
sky mountain snow clouds water tree
sky mountain snow water tree

field foals horses mare
horses field foals mare people flowers tree
horses foals mare field grass tree
horses foals mare field grass

people pillars stone temple
stone pillar sculpture statue tree temple people
pillars sculpture statue stone temple people
pillars sculpture statue stone temple people

**Figure 1: Annotation examples. First line is the annotation from Corel, second is LSA1, third is NLSA and last is NLSA+IWE. Except in LSA1, the keywords order in inference-based annotation methods is defined by the posterior probabilities rank.**

**Table 1: Comparative performance: maximum normalized score vs. number of latent components $K$ for all latent space models. LSA1 represents LSA with annotation propagation, LSA2 represents LSA with soft inference, NSA+IWE denotes NSA followed by IWE-based inference.**

| Method | Number of Latent Components: $K$ | | | | |
|---|---|---|---|---|---|
| | 20 | 40 | 60 | 80 | 100 |
| PLSA | 0.445 | 0.447 | 0.450 | 0.448 | 0.446 |
| LSA1 | 0.490 | 0.517 | 0.521 | 0.536 | 0.546 |
| LSA2 | 0.493 | 0.513 | 0.523 | 0.535 | 0.548 |
| NLSA | 0.535 | 0.567 | 0.583 | 0.602 | 0.598 |
| NLSA+IWE | 0.548 | 0.583 | 0.608 | 0.624 | 0.615 |

CDs, of which each consists of 5000 training images and 2000 testing images in average. The average textual vocabulary size per subset is 150, and the average textual keyword number for each annotated image is 3. Annotated or unannotated image feature representation has been clearly described in Section 2. The kernel function $k(\mathbf{x}, \mathbf{y})$ involved in NLSA is defined as the Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = \exp\{-\|\mathbf{x} - \mathbf{y}\|^2/2\sigma^2\}$, and the value of $Z$ involved in top-$Z$ rank is set to 3. The dimension of the embedding space $\{\mathbf{r}_n\}$ is predefined as 2, which facilitates fast learning and inference processes that IWE requires.

To evaluate annotation accuracy on a dataset with the vocalulary size $L$, we use the normalized score measure [1] $E_{NS}^{method} = r/l - w/(L - l)$, where $l$ denotes the actual number of keywords in the test image and $r$ is the number of correctly predicted words, $w$ denotes the wrongly predicted number of keywords on the contrary. This measure can be used for any of the annotation procedures described in this paper.

For PLSA [8], LSA followed by soft inference, NLSA, and IWE-based inference after NLSA, the normalized score varies according to a variational threshold level. For LSA with propagation [8], no probability is attached to each ranked keyword, hence the threshold level cannot be applied directly. The way to deal with it is to compute the average number of predicted words at each threshold level over all subsets, the corresponding normalized score is then computed subject to the number. By tuning the threshold level, we report the corresponding maximum normalized scores under different number $K$, i.e. the number of latent space components (aspects), in Table 1. We conclude that PLSA is the baseline annotation method, LSA1 and LSA2 are very close in performance, while a larger improvement is observed for NLSA and NLSA+IWE, with the latter as the best annotation approach among comparative annotation procedures. Some real image auto-annotation examples are shown in Figure 1.

## 7. CONCLUSIONS

We proposed a new unsupervised image auto-annotation system, which comprises two serial annotation procedures NLSA and IWE. First, NLSA trains a nonlinear mapping, which spans the nonlinear latent space. Later soft inference is introduced to attach probability to each ranked keyword. Based on the inferred posterior distribution over the keyword vocabulary, IWE re-infers the posteriors via modeling the interacted nature inherent in multi-modal data.

## 8. ACKNOWLEDGMENTS

The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region. The work was conducted at The Chinese University of Hong Kong.

## 9. REFERENCES

[1] K. Barnard, P. Duygulu, N. Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107-1135, 2003.

[2] D. Blei and M. I. Jordan. Modeling annotated data. In *Proc. SIGIR*, Toronto, Canada, July 2003.

[3] E. Chang, G. Kingshy, G. Sychay, and G. Wu. CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines. *IEEE Trans. on CSVT*, 13(1):26-38, Jan. 2003.

[4] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391-407, 1990.

[5] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177-196, 2001.

[6] T. Iwata, K. Saito, N. Ueda, S. Stromsten, T. Griffiths, and J. Tenenbaum. Parametric Embedding for Class Visualization. In *Proc. of NIPS*, Vancouver, Canada, Dec. 2004.

[7] 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.

[8] F. Monay and D. Gatica-Perez. On Image Auto-Annotation with Latent Space Models. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, Berkeley, California, USA, Nov. 2003.

[9] F. Monay and D. Gatica-Perez. PLSA-based Image Auto-Annotation: Constraining the Latent Space. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, New York, USA, Oct. 2004.

[10] B. Schölkopf, A. Smola, and K. R. Müller. Nonlinear Component Analysis as A Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299-1319, 1998.