# On Image Auto-Annotation with Latent Space Models

Florent Monay
Dalle Molle Institute for
Perceptual Artificial Intelligence (IDIAP)
Martigny, Switzerland
monay@idiap.ch

Daniel Gatica-Perez
Dalle Molle Institute for
Perceptual Artificial Intelligence (IDIAP)
Martigny, Switzerland
gatica@idiap.ch

## ABSTRACT

Image auto-annotation, i.e., the association of words to whole images, has attracted considerable attention. In particular, unsupervised, probabilistic latent variable models of text and image features have shown encouraging results, but their performance with respect to other approaches remains unknown. In this paper, we apply and compare two simple latent space models commonly used in text analysis, namely Latent Semantic Analysis (LSA) and Probabilistic LSA (PLSA). Annotation strategies for each model are discussed. Remarkably, we found that, on a 8000-image dataset, a classic LSA model defined on keywords and a very basic image representation performed as well as much more complex, state-of-the-art methods. Furthermore, non-probabilistic methods (LSA and direct image matching) outperformed PLSA on the same dataset.

## Categories and Subject Descriptors

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

## General Terms

Algorithms, Theory

## Keywords

automatic annotation of images, semantic indexing

## 1. INTRODUCTION

Searching image collections is intuitive when adequate annotations are available. Words are inherently semantic, and standard keyword-based search techniques can efficiently compute similarities between text-based queries and image captions, satisfying the requirements of many image users. Of course, images have to be first annotated, but most of them are not labeled at production time, and off-line annotation is laborious and expensive. It is hence not suprising that image auto-annotation has attracted attention in the literature [2, 1, 4, 7, 3]. In this rich domain, we refer to annotations as nouns that describe the image content, i.e., objects (e.g. "mountain") or concepts (e.g., "sunset").

Image auto-annotation has been addressed from two different perspectives. The first one defines annotation as a supervised learning problem, and associates words to images by first defining classes, each one corresponding to a word [4], or a set of words defining a concept [7], followed by training of each visual class model with manually labeled images, image classification into one or more classes, and finally annotating by *propagating* the corresponding class words. This approach clearly separates the textual from the visual components, computing similarity at the visual level.

The second approach takes a differing viewpoint, and attempts to discover the statistical links between visual features and words on an unsupervised basis, by estimating the joint distribution of words and regional image features, and elegantly posing annotation as statistical *inference* in a graphical model [1]. The proposed joint models account for the distinct data nature, and do not need labeled data. Further work has also investigated region naming, i.e., the association of words to specific image regions [1, 3].

Given the recent emergence of this field, no common corpora, evaluation measures and protocols have been defined. Furthermore, objectively assessing the quality of image auto-annotation is in itself a complex problem [1, 7]. While it is not possible to derive a direct comparison between current algorithms (most of which are complex), several questions of interest remain open. First, how superior are state-of-the-art methods compared to simpler approaches? Second, is annotation by propagation better than annotation by inference in practice? Third, how well do methods scale up?

This paper addresses the first two questions for the case of unsupervised methods. We apply and compare two well-known latent space models for discrete data: LSA [6] and PLSA [8]. Our work advocates for a systematic, comparative evaluation of algorithms using common measures and datasets, acknowledging the difficulty of the annotation evaluation tasks, and describes two interesting and somewhat surprising empirical results. First, we show that a very simple approach performs as well as state-of-the art methods [1]. Second, we show that non-probabilistic methods based on annotation by propagation (LSA and direct image matching) outperformed the probabilistic formulation, which performs annotation by inference.

The paper is organized as follows. Section 2 describes the representation of annotated images as discrete data. Section 3 discusses LSA and PLSA. Section 4 describes their appli-


*MM'03,* November 2–8, 2003, Berkeley, California, USA.

cation to auto-annotation. Section 5 presents the results and discussion. Section 6 concludes the paper.

## 2. ANNOTATED IMAGES IN A VECTOR SPACE REPRESENTATION

Several models for collections of discrete data have been proposed and successfully applied to text analysis [6, 8]. In such models, each document $d_i$ in a collection consists of a set of words, and is represented in a simple vector format, where the j-th vector component is the frequency of $w_j$, the j-th word in the vocabulary. A text corpus is hence summarized by a *term-by-document* matrix $A \in \mathbb{R}^{N*M}$, where $N$ is the number of documents and $M$ is the vocabulary size.

Annotated images (multimedia documents) can naturally be embedded in such a vector-space representation in order to apply text analysis methods, via a quantized image representation [9, 10]. In this paper, we use a very simple one. Images are first segmented into three fixed regions that comprise the image center, and the upper and lower parts (Fig. 1) (professional images like Corel's often depict the main objects in their center). For each region, a $6 \times 6 \times 6$ RGB color histogram is computed, leaving an image feature vocabulary of 648 terms. More elaborate features could be added in a straightforward fashion [1, 4, 7].

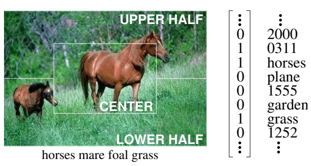


**Figure 1: From annotated image to vector-space. The four digits numbers are the *visual keywords*: the first digit encodes the region number and the last three digits are the coordinates of the 6x6x6 RGB value.**

Annotated images are modeled by concatenated feature vectors of word and image features (Fig. 1). When a distinction in needed, a *keyword* will refer to the words, and *visual keyword* to the visual features. More generally, we refer to keywords and visual keywords as *terms*, and the vocabulary is therefore the set of all observed terms in a dataset. Non-annotated images are represented in the full vocabulary vector space, with all elements corresponding to keywords set to zero.

## 3. LATENT SPACE MODELS

Two documents can be similar from a semantic viewpoint even if their words or visual features are not identical: different words can be used to express the same concept (synonymy), and several colors can represent the same object. Furthermore, the same word (or color) might have different meanings depending on the context (polysemy). Modeling directly at the word or visual feature level would miss these ambiguities [6]. Existing approaches are based on the definition of a *latent space* where the documents are represented in a disambiguated form. Latent Semantic Analysis (LSA) [6] and Probabilistic Latent Semantic Analysis (PLSA) [8] are two such algorithms, investigated here in the context of auto-annotation.

### 3.1 LSA

A classic algorithm arising from linear algebra, LSA decomposes the term-by-document matrix in three matrices by a truncated Singular Value Decomposition (SVD),

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

where $A \in \mathbb{R}^{N*M}$, $U \in \mathbb{R}^{N*K}$, $S \in \mathbb{R}^{K*K}$ and $V \in \mathbb{R}^{M*K}$. 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. LSA has been extensively used in text analysis, and more recently to improve retrieval of multimedia news documents [9, 10]. Unfortunately, LSA lacks a clear probabilistic interpretation [8].

### 3.2 PLSA

PLSA [8] models each term in a document as arising from a mixture model. The mixture components are multinomial latent variables that represent *aspects* or *topics*. A word can come from more than one aspect, and documents can therefore contain multiple aspects. In this model, each observed term $w_j$ is conditionally independent of the document $d_i$ it belongs to given an aspect variable $z_k$. The term-document joint probability, assuming $K$ aspects, is given by:

$$P(w_j, d_i) = P(d_i) \sum_{k=1}^{K} P(w_j \mid z_k) P(z_k \mid d_i). \tag{2}$$

As usual, maximum likelihood parameter estimation is performed with the Expectation-Maximization (EM) algorithm. The large number of parameters in the model ($M * K$ for $P(w_j \mid z_k)$ and $K * N$ for $P(z_k \mid d_i)$) makes PLSA prone to overfitting, and requires a *tempered* version of EM [8].

## 4. IMAGE AUTO-ANNOTATION

We now discuss three studied annotation strategies. The first two are based on comparison and annotation propagation. The third one is based on statistical inference.

### 4.1 Annotation by direct match

The simplest method consists of two steps: (i) similarity computation in the vector space between the image to annotate and each image in the annotated corpus, using a standard cosine measure, and (ii) keyword propagation from the corpus on an image-by-image basis, depending on the similarity rank.

### 4.2 Annotation with LSA

Once a document collection has been processed (section 3.1), the similarity between an unannotated image $q \in \mathbb{R}^{1*M}$ and the annotated image corpus is measured in the latent space. $q$ is first projected by right multiplying by V, the terms expressed in the latent space basis,

$$\hat{q} = q * V.$$

After projection, the similarity between $\hat{q}$ and each row of $U$ (representation of the collection in the latent space) is computed using the cosine measure. The annotation is then propagated from the ranked documents. Annotations are less reliable as the similarity between documents decreases.

### 4.3 Annotation with PLSA

Unlike LSA, PLSA allows us to define annotation as a process of computing probabilities, in particular, the posterior distribution of the terms of the vocabulary given an unannotated image $q$. From Eq. 2,

$$P(w_j \mid q) = \sum_{k=1}^{K} P(w_j \mid z_k) P(z_k \mid q).$$

For annotation, the distributions $P(w_j \mid z_k)$ are estimated once from the training set, while the $P(z_k \mid q)$ topic mixture for each unannotated image $q$ is computed following the procedure described in [8] [1]. The posterior distributions over keywords are then selected and renormalized, which creates a soft annotation over the full keyword vocabulary.

# 5. RESULTS AND DISCUSSION

## 5.1 Dataset

Large public annotated image datasets are not common. Different subsets of the Corel image collection (60000 images annotated with 3 to 5 describing keywords) have been used in recent work [1, 7, 4]. However, no common samples have been defined. Barnard et. al. [1] presented a study of different auto-annotation methods on a subset of 80 Corel CDs, from which 10 different training and test sets were sampled. The average number of images for training is 5200 and 1800 for testing. Performance evaluation is especially well addressed in their work, and therefore we have used a similar dataset for comparison. We recreated nine of the 10 datasets, with more than 98% intersection with [1]. Furthermore, as in [1], the keyword vocabulary size was reduced from an average of 1876 to 149 keywords, by retaining only the keywords appearing more than 20 times in the training set. The resulting empirical keyword distributions for one of the nine training subsets is shown in Fig. 2.

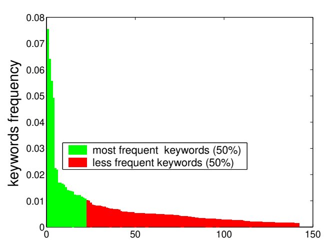


**Figure 2: Empirical keyword distribution in sample set 1. The 23 most common keywords, accounting for 50% of the probability mass are (in order): water, sky, people, trees, grass, clouds, bird, snow, stone, street, building, jet, pattern, buildings, texture, tree, plane, fish, coast, rocks, mountains, beach, and ground.**

## 5.2 Performance Evaluation

Many different measures can be considered to evaluate the annotation accuracy of an algorithm, but some important points have to be considered. First, annotations in test data might not include some "correct" keywords. For comparative purposes, this does not represent a problem as all algorithms have to deal with this issue, but the estimated measures can indeed be over-pessimistic. Second, the vocabulary statistics (Fig. 2) must be taken into account, because predicting very frequent words like "water" or "people" are safer guesses than less frequent words. An automatic annotation method should therefore perform better than simply using the empirical word distribution of the training set. Third, the number of correctly predicted words $r$ for an image has to be somewhat penalized by an increasing number of wrong predicted words $w$ with respect to the vocabulary size $N$. We use the *normalized score* measure [1],

$$E_{NS}^{method} = r/n - w/(N - n),$$

where $n$ denotes the actual number of keywords in the test image. This measure is related to precision and recall. Predicting exactly the right $n$ keywords implies $E_{NS} = 1$, predicting all but the right $n$ keywords produces a value of -1, and predicting all the vocabulary produces a zero value.

This measure can be used for any of the annotation procedures described in Section 4. For PLSA and empirical words distribution, the normalized score is plotted by varying a threshold level and predicting the words with a posterior probability higher than this *refuse to predict level.* For LSA and direct match, no probability is attached to each ranked keyword, hence the threshold level cannot be applied directly. To overcome this problem, we first compute the average number of predicted words for PLSA at each threshold level over all the nine subsets. The normalized scores for LSA and direct match methods are then computed at each threshold level for the corresponding average number of predicted words.

## 5.3 Experimental Results

For our experiments we limited the number of visual keywords for each image by an empirical threshold of 0.04 on the normalized RGB histograms: visual keywords with a lower probability are not attached to the image. This leads to an average of 18.5 visual keywords per image, which is a trade-off between keeping enough visual information and balancing the amount of visual and textual keywords (the latter with an average number of 3.6). Given this threshold, 525 visual keywords are present on average in each sampled vocabulary.

For the aspect models, we varied the number of aspects from 15 to 80, and reported the corresponding maximum normalized score in Table 1. PLSA results barely changed with varying number of aspects, while a larger improvement was observed when increasing the number of aspects for LSA. We tried to initialize the EM training prodedure in PLSA with probabilities derived from the LSA decomposition of the training set[2], but this did not improve the perplexity and the annotation performance over the standard random initialization. In the rest of the paper, the results correspond to 60 latent aspects for both PLSA- and LSA-based annotation methods[3].

| Method | Number of aspects $K$ | | | |
|---|---|---|---|---|
| | 15 | 40 | 60 | 80 |
| LSA | 0.495 | 0.526 | 0.531 | 0.535 |
| PLSA | 0.447 | 0.449 | 0.452 | 0.446 |

**Table 1: Maximum normalized score vs. number of latent aspects $K$ for PLSA and LSA.**

Figure 3 shows the normalized score of five annotation methods: usage of the empirical word distribution (emp.),

[1] PLSA is not a fully generative model. See [3] for discussion.

[2] An empirical method to derive probabilities from LSA is suggested in [5].

[3] More results will be made available in an extended version of this paper.

propagation after direct match in the original feature space (dir.), propagation after LSA on all the terms (LSA 1), propagation after LSA on visual features only (LSA 2), and computation of the posterior probability of each keyword given the unannotated image by PLSA (PLSA). For all the the five methods, the average number of predicted words corresponding to the best annotation performance (highest Normalized Score) is around 40.

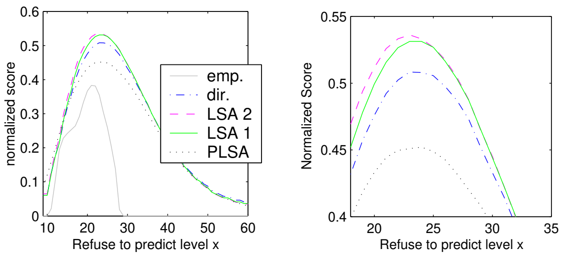


**Figure 3: Normalized score for all the methods vs. refuse to predict level $x$ ($p = 10^{-x/10}$ where $p$ is the probability threshold).**

The maximum normalized score increases over the empirical distribution for our methods is shown in Table 2. For comparison with the state-of-the-art, the results reported in [1] range from 0.107 for *MoM-LDA* model to 0.179 for *binary-I-2-region-cluster* and *binary-I-2-region-cluster*; typical increase is 0.160. It is interesting to notice that even with very basic image features such as the one used in this paper, standard methods such as LSA on image features only (LSA 2) can achieve similar annotation results compared to complex, fully generative probabilistic models.

| Measure | Method | | | |
|---|---|---|---|---|
| | dir. | LSA 1 | LSA 2 | PLSA |
| $diff$ | 0.125 | 0.148 | 0.153 | 0.069 |

**Table 2: Difference between the maximum normalized score of the empirical distribution (0.383) and the maximum of each discussed methods ($diff = max(E_{NS}^{method}) - max(E_{NS}^{emp})$).**

The fact that the PLSA annotation score is lower than LSA is somewhat contradictory with the results presented in [8], where PLSA outerperforms LSA for retrieving text. Several reasons could explain this difference. One possible reason could be that propagating annotation can lead to good results especially when annotation is uniform in a given subset: if some images are systematically annotated with the same set of words, propagation methods can find the exact annotation if the right image is retrieved. This phenomena is illustrated in Figure 4. On the first and second images, direct match and LSA methods have retrieved an example with a very similar annotation to propagate the keywords from, thus finding a highly accurate annotation. PLSA, which attempts to model complete distributions, can find a relevant annotation for the last image but annotates the second image with completely off-topic words. On the first image, PLSA provides the words *grass trees* and *sky*, which are not in the original annotation from Corel but could be appropriately attached to the image. This ability of PLSA to handle polysemy [8] could be penalized by the way of evaluating annotation.


**Figure 4: Annotation examples with four keywords: first line is the annotation from Corel, second is direct match, third is LSA 1 and last is PLSA. The keywords order is defined by the original Corel annotation for propagation-based methods (direct match and LSA). PLSA annotation is ordered by posterior probabilities. The empirical word distribution annotation is the same for all images: *water sky people trees.***

# 6. CONCLUSION AND FUTURE WORK

We tested and evaluated two latent space models on a very basic representation of annotated images. The performance of auto-annotation derived from some of these simple models were comparable to much more complex methods on a 8000-images dataset. Annotation by propagation (LSA and direct match) outperformed annotation by inference (PLSA), suggesting that propagation is a good strategy for that type of dataset and vocabulary size (∼150 keywords). The methods performance on a larger vocabulary remains an open question that will be addressed in the future.

## Acknowledgements

We thank Thomas Hofmann (code), Kobus Barnard and James Z. Wang (data) for their help. This work was funded by the swiss NCCR on Interactive Multimodal Information Management (IM)2. The images used in this study belong to the Corel stock photo collection ©.

# 7. 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] A. Benitez and S.-.F. Chang. Semantic knowledge construction from annotated image collections. In *Proc. IEEE ICME*, Lausanne, Jul. 2002.

[3] D. Blei and M. I. Jordan. Modeling annotated data. In *Proc. SIGIR*, Toronto, Aug. 2003.

[4] 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.

[5] N. Coccaro and D. Jurafsky. Towards better integration of semantic predictors in statistical language modeling, In *Proc. ICSLP*, Sydney, Nov. 1998.

[6] 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.

[7] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on PAMI*, 25(10), Oct. 2003.

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

[9] T. Westerveld. Image retrieval: Content versus context. In *Proc. RIAO*, Paris, 2000.

[10] R. Zhao and W. Grosky. Narrowing the semantic gap - improved text-based web document retrieval using visual features. *IEEE Trans. on Multim.*, 4(2):189-200, Jun. 2002.