# PLSA-based Image Auto-Annotation: Constraining the Latent Space

Florent Monay
monay@idiap.ch

Daniel Gatica-Perez
gatica@idiap.ch

IDIAP Research Institute
Rue du Simplon 4, CP 592
1920 Martigny, Switzerland

## ABSTRACT

We address the problem of unsupervised image auto-annotation with probabilistic latent space models. Unlike most previous works, which build latent space representations assuming equal relevance for the text and visual modalities, we propose a new way of modeling multi-modal co-occurrences, constraining the definition of the latent space to ensure its consistency in semantic terms (words), while retaining the ability to jointly model visual information. The concept is implemented by a linked pair of Probabilistic Latent Semantic Analysis (PLSA) models. On a 16000-image collection, we show with extensive experiments that our approach significantly outperforms previous joint models.

## Categories and Subject Descriptors

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

## General Terms

Algorithms, Theory, Languages

## Keywords

Automatic Annotation of Images, Semantic Indexing, PLSA

## 1. INTRODUCTION

The potential value of large image collections can be fully realized only when effective methods for access and search exist. Image users often prefer to formulate intuitive text-based queries to retrieve relevant images [1], which requires the annotation of each image in the collection. Automatic image annotation has thus emerged as one of the key research areas in multimedia information retrieval [3, 4, 2], as an alternative to costly, labor-intensive manual captioning.

Motivated by the success of latent space models in text analysis, generative probabilistic models for auto-annotation have been proposed, including variations of PLSA [5], and Latent Dirichlet Allocation (LDA) [2]. Such models use a latent variable representation for unsupervised learning of co-occurrences between image features and words in an annotated image collection, and later employ the learned models to predict words for unlabeled images [4, 2, 6]. The latent space representation can capture high-level relations within and across the textual and visual modalities.

Specific assumptions introduce variations in the ways in which co-occurrence information is captured. However, with a few exceptions [2], most previous works assume that words and visual features should have the same importance in defining the latent space [4, 6]. There are limitations with this view. First, the semantic level of words is much higher than the one of visual features extracted even by state-of-the-art methods. Second, in practice, visual feature co-occurrences across images often do not imply a semantic relation between them. This results in a severe degree of visual ambiguity that in general cannot be well handled by existing joint models. For auto-annotation, we are ultimately interested in defining a latent space that is consistent in semantic terms, while able to capture multimodal co-occurrences.

We present a novel approach to achieve the above goal, based on a linked pair of PLSA models. We constrain the definition of the latent space by focusing on textual features first, and then learning visual variations conditioned on the space learned from text. Our model consistently outperforms previous latent space models [6], while retaining the elegant formulation of annotation as probabilistic inference.

The paper is organized as follows. Section 2 describes our representation of annotated images. Section 3 presents the key PLSA concepts. Section 4 introduces our approach, motivated by the limitations of previous models. Section 5 presents experiments. Section 6 concludes the paper.

## 2. DATA REPRESENTATION

Annotated images are *documents* combining two complementary modalities, each one referring to the other: while an image potentially illustrates hundreds of words, its caption specifies the context. Both textual and visual modalities are represented in a discrete *vector-space* form.

**Caption.** The set of captions of an annotated image collection defines a *keywords vector-space* of dimension $W$, where each component indexes a particular keyword $w$ that occurs in an image caption. The textual modality of a particular document $d$ is thus represented as a vector $t_d = (t_{d1}, \ldots, t_{dw}, \ldots, t_{dW})$ of size $W$, where each element $t_w$ is the count of the corresponding word $w$ in document $d$.

**Image.** We use two common image representations.

*RGB* [6]: 6∗6∗6 RGB histograms are computed from three distinct regions in the image, and only values higher than a threshold value are kept. This amounts at keeping only the dominant colors. The RGB vector-space is then built from the bin values found in the whole image set with respect to the three regions. The visual modality of document $d$ is then $v_d = (v_{d1}, \ldots, v_{db}, \ldots, v_{dB})$, a vector of size $B = 6^3 * 3$.

*Blobs* [3] : The normalized cut segmentation algorithm is applied to the image set, and the resulting regions are represented by color, texture, shape, size, and position descriptors. The K-means clustering algorithm is applied to all the computed descriptors, quantizing the image regions into a $B$-dimensional *blob vector-space* (same notation as RGB).

# 3. THE PLSA MODEL

In a collection of discrete data such as the annotated image dataset described in Section 2, a fundamental problem might occur: different elements from the vector-space can express the same concept (*synonymy*) and one element might have different meanings depending on the context (*polysemy*). If this semantic issue is well known for text, visual data share similar ambiguities: one color might have different meanings if occurring with different sets of color and two colors could represent the same concept.

When this ambiguities occur, a disambiguate *latent space* representation could potentially be extracted from the data, which is the goal of PLSA [5]. This model assumes the existence of a latent variable $z$ (aspect) in the generative process of each element $x_j$ in a particular document $d_i$. Given this unobserved variable, each occurence $x_j$ is independent from the document it was generated from, which corresponds to the following joint probability: $P(x_j, z_k, d_i) = P(d_i)P(z_k \mid d_i)P(x_j \mid z_k)$. The joint probability of the observed variables is obtained by marginalization over the $K$ latent aspects $z_k$,

$$P(x_j, d_i) = P(d_i) \sum_{k}^{K} P(z_k \mid d_i) P(x_j \mid z_k). \qquad (1)$$

**Model parameters**. The PLSA parameters are the two conditional distributions in equation 1, and are computed by an Expectation-Maximization algorithm on a set of training documents [5]. For a vector-space representation of size $N$, $P(x \mid z)$ is a $N$-by-$K$ table that stores the parameters of the $K$ multinomial distributions $P(x \mid z_k)$. To give an intuition of $P(x \mid z)$, Figure 3 (b) shows the posterior distribution of the 10 most probable words for a given aspect, for a model trained on a set of image captions. The keywords distribution refers to a *people and costume*-related set of keywords. $P(x \mid z)$ characterizes the aspect, and is valid for documents out of the training set [5].

On the contrary, the other $K$-by-$M$ table $P(z \mid d)$ is only relative to the $M$ training documents. Storing the parameters of the $M$ multinomial distributions $P(z \mid d_i)$, it does not carry any a priori information about the probability of aspect $z_k$ beeing expressed in any unseen document.

**Learning**. The standard Expectation-Maximization approach is used to compute the model parameters $P(x \mid z)$ and $P(z \mid d)$ by maximizing the data likelihood.

$$\mathcal{L} = \prod_{i}^{M} \prod_{j}^{N} P(d_i) \sum_{k}^{K} P(z_k \mid d_i) P(x_j \mid z_k)^{n(d_i, x_j)}, \qquad (2)$$

where $n(d_i, x_j)$ is the count of element $x_j$ in document $d_i$.

**E-step :** $P(z \mid d, x)$, the probabilities of latent aspects given the observations are computed from the previous estimate of the model parameters (randomly initialized).

**M-step :** The parameters $P(x \mid z)$ and $P(z \mid d)$ are updated with the new expected values $P(z \mid d, x)$.

**Inference: PLSA of a new document**. For an unseen document $d_{new}$, the conditional distribution over aspects $P(z \mid d_{new})$ has to be computed. The method proposed in [5] consist in maximizing the likelihood of the document $d_{new}$ with a partial version of the EM algorithm described above, where $P(x \mid z)$ is *kept fixed* (not updated at each M-step). In doing so, $P(z \mid d_{new})$ maximizes the likelihood of document $d_{new}$ with respect to the previously trained $P(x \mid z)$ parameters.

# 4. PLSA-BASED ANNOTATION

PLSA has been recently proposed as a model for automatic image annotation [6]. Referred here as PLSA-MIXED, it somewhat showed surprisingly poor annotation performance with respect to very basic non probabilistic methods [6]. We propose here a new application of PLSA to automatic image annotation and motivate our approach by an analysis of PLSA-MIXED, which then leads to the new method.

## 4.1 PLSA-mixed

The PLSA-MIXED system applies a standard PLSA on a *concatenated representation* of the textual and the visual modalities of a set of annotated images $d$: $x_d = (t_d, v_d)$. Using a training set of captioned images, $P(x \mid z)$ is learned for both textual and visual co-occurrences, which is an attempt to capture simultaneous occurrence of visual features (regions or dominant colors) and words. Once $P(x \mid z)$ has been learned, those parameters can be used for the auto-annotation of a new image.

The new image $d_{new}$ is represented in the concatenated vector space, where all keywords elements are zero (no annotation): $x_{new} = (0, v_{new})$. The multinomial distribution over aspects given the new image $P(z \mid d_{new})$ is then computed with the partial PLSA steps described in Section 3, and allows the computation of $P(x \mid d_{new})$. From $P(x \mid d_{new})$, the marginal distribution over the keyword vector-space $P(t \mid d_{new})$ is easily extracted. The annotation of $d_{new}$ results from this distribution, either by selecting a predefined number of the most probable keywords or by thresholding the distribution $P(t \mid d_{new})$.

## 4.2 Problems with PLSA-mixed

Using a concatenated representation, PLSA-MIXED attempts to simultaneously model visual and textual modalities with PLSA. It means that intrinsically, PLSA-MIXED assumes that the two modalities have an equivalent importance in defining the latent space. This has traditionally been the assumption in most previous work [4]. However, an analysis of the captions and the image features in the Corel dataset (described in Section 5) emphasizes the difference between the keywords and the visual features occurrences. Figure 1 shows two similarity matrices for a set of annotated images ordered by topics, as in human-based CD organization provided by Corel. They represent the cosine similarity between each document in the keyword space (left), and the visual feature space (Right). The keywords similarity matrix has sharp block-diagonal structure, each corresponding to a consistent cluster of images, while the second similarity matrix (visual features) consist in a less contrasted pattern.

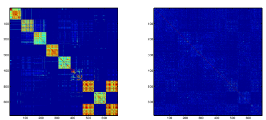

**Figure 1: Similarity matrices for a set of manually ordered documents (9 CDs from Corel). The left matrix is the textual modality, the right matrix is the visual modality (Blobs features are used).**

Of course, Figure 1 does not prove that no latent representation exists for the visual features, but it strongly suggests that in general, two PLSA separately applied on each modality would define two distinct latent representations of the same document. For example, color co-occurrence happens across images, but does not necessarily mean that the corresponding images are semantically related. PLSA-MIXED thus might model aspects mainly based on visual features, which results in a prediction of almost random keywords if these aspects have high probabilities given the image to annotate. Moreover, assuming that no particular importance is given to any modality, the amount of visual and textual information need to be balanced in the concatenate representation of an annotated image. This constrains the size of the visual representation, as the number of keywords per image is usually limited (an average of 3 for the data we used). A typical aspect from PLSA-MIXED where images are relatively consistent in terms of visual features, but not semantically (dominant colors: green, red, yellow, black) is shown in Figure 2.

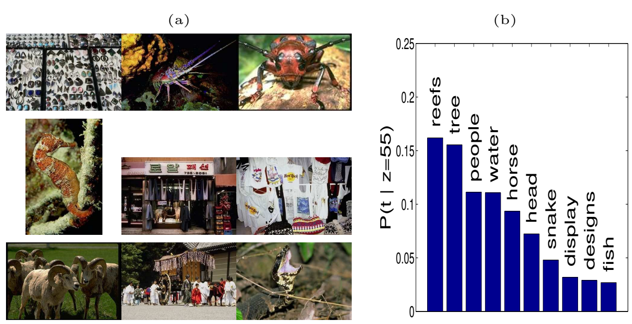


**Figure 2: One semantically meaningless aspect from PLSA-MIXED: the 9 most probable images in the training set and the 10 most probable keywords with their corresponding probability $P(t \mid z)$.**

## 4.3 Our approach: PLSA-words

Given the above observations, we propose to model a set of documents $d$ with two linked PLSA models sharing the same distribution over aspects $P(z \mid d)$. Contrarily to PLSA-MIXED, this formulation allows to treat each modality differently and give more importance to the captions in the latent space definition. The idea is to capture meaningful aspects in the data and use those for annotation. Both parameters estimation and annotation inference involve two linked PLSA steps[1].

[1]Computational complexity is discussed at www.idiap.ch/∼monay/acmm04/

**Learning parameters**

**1**. A first PLSA model is completely trained on the set of image captions to learn both $P(t \mid z)$ and $P(z \mid d)$ parameters. Figure 3 illustrates one aspect automatically learned on the textual modality, with its most probable training images (a ) and their corresponding distribution over keywords $P(t \mid z)$ (b ). This example[2] shows that this first PLSA can capture meaningful aspects from the data.

**2** . We then consider that the aspects have been observed for this set of documents $d$ and train a second PLSA on the visual modality to compute $P(v \mid z)$, keeping $P(z \mid d)$ from above fixed. Note that this technique is very similar to the process described in Section 3, where $P(x \mid z)$ was kept fixed and $P(z \mid d)$ was computed by likelihood maximization.

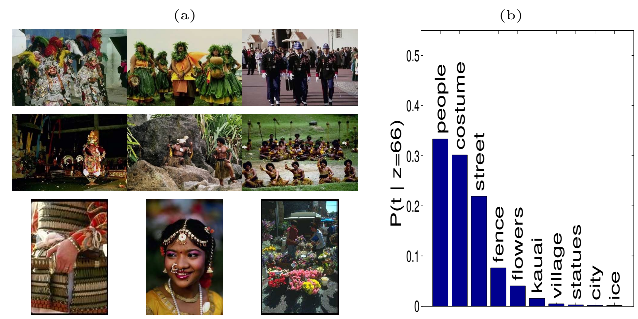


**Figure 3: One aspect from PLSA learned on words: the 9 most probable images in the training set (from $P(z \mid d)$) and the 10 most probable keywords with their corresponding probability $P(t \mid z)$.**

**Annotation by inference**

**1**. Given new visual features $v_{new}$ and the previously calculated $P(v \mid z)$ parameters, $P(z \mid d_{new})$ is computed for a new image $d_{new}$ using the standard PLSA procedure for a new document (Section 3).

**2**. The posterior probability of keywords given this new image is then inferred by:

$$P(t \mid d_{new}) = \sum_{k}^{K} P(t \mid z_k) * P(z_k \mid d_{new}) \quad (3)$$

If a new image has a high probability of belonging to one aspect, then a consistent set of keywords will be predicted. The PLSA-WORDS method thus automatically builds a kind of *language model* for the set of training images, which is then applied for auto-annotation. It is also interesting to notice that PLSA is applied here on very small textual documents, given that each annotation is about 3 words long.

# 5. PERFORMANCE EVALUATION

## 5.1 Data

The data used for experiments are comprised of roughly 16000 Corel images split in 10 overlapping subsets, each devided in training (∼5200 images) and testing sets (∼1800 images). The average vocabulary size per subset is 150 keywords, and the average caption size is 3. Both RGB and Blobs features described in Section 2 are tested. Blob features were downloaded from Kobus Barnard's website [4].

[2]Find more examples at www.idiap.ch/∼monay/acmm04/

## 5.2 Performance measures

No commonly agreed image auto-annotation measure exists. We evaluated our method on three different measures, but restrict the discussion to the two measures described below for space reasons[3].

**Annotation accuracy :** When predicting exactly the same number of keywords as the ground truth, the *annotation accuracy* for one image is defined as $Acc = r/n$, where $r$ is the number of correctly predicted keywords and $n$ is the size of the ground truth caption. The average annotation accuracy is computed over a set of images.

**Normalized Score [4] :** Sharing the same $r$ and $n$ values with the above definition, the normalized score is defined as: $Nscore = r/n-(p-r)/(N-n)$, where $N$ is the vocabulary size and $p$ is the number of predicted keywords. The average normalized score is computed over a set of images for a varying number of predicted keywords and the maximum is reported here.

## 5.3 Results

We compare the two PLSA-based methods described in Section 4.1 and 4.3, and three other methods : EMPIRICAL, LSA and PLSA-SPLIT. EMPIRICAL simply uses the empirical keywords distribution from the training set to predict the same set of keywords regardless of the image content; LSA was the best method reported in [6] in term of normalized score, better than PLSA-MIXED; and PLSA-SPLIT is the *unlinked* equivalent of PLSA-WORDS, for which two distinct sets of parameters $P_t(z \mid d)$ and $P_v(z \mid d)$ are learned for each modality. The latent space dimensionality $K = 100$ has been used for all the reported results (except EMPIRICAL). The average annotation accuracy results are presented in Table 1 and Table 2 contains the maximum normalized scores values. All results are averaged over the 10 subsets.

| Method | BLOBS | RGB |
|---|---|---|
| EMPIRICAL | 0.191 (0.012) | 0.191 (0.012) |
| LSA | 0.140 (0.009) | 0.178 (0.009) |
| PLSA-SPLIT | 0.113 (0.017) | 0.121 (0.019) |
| PLSA-MIXED | 0.221 (0.011) | 0.217 (0.024) |
| PLSA-WORDS | 0.292 (0.011) | 0.288 (0.014) |

**Table 1: Average annotation accuracy computed over the 10 subsets. These values correspond to an average number of 3.1 predicted keywords per image. The variance is given in parantheses.**

The RGB and Blobs features give similar annotation performance for both measures. This suggests that the blob representation is equivalent to the much simpler RGB features when applied to this annotation task. One explanation could be that the k-means algorithm applied on the concatenated color and texture representation of the image regions converges to a color-only driven clustering.

As originally reported [6], the PLSA-MIXED maximum normalized score is lower than the non-probabilistic LSA one, while PLSA shows better performance than LSA for textual data modeling [5]. Annotation accuracy, which measures the quality of smaller but more realistic annotation, gives PLSA-MIXED as the best performing method.

The ranking of the three PLSA-based methods emphasizes the importance of a well defined link between textual and visual modalities. PLSA-SPLIT naively assumes no link between captions and images and models them separately. No match between the two latent space definitions exist, which explains why PLSA-SPLIT performs worse than the simplest EMPIRICAL method. The PLSA-MIXED method introduces a determining yet unclear interaction between text and image by concatenating the two modalities. This connexion translates in significant improvement over PLSA-SPLIT in both annotation and normalized score measures.

PLSA-WORDS outperforms both PLSA-SPLIT and PLSA-MIXED, therefore justifing its design. PLSA-WORDS makes an explicit link between visual features and keywords, learning the latent aspects distribution in the keywords space and fixing these parameters to learn the distribution of visual features. This results in the definition of semantically meaningfull clusters, and forces the system to predict consistent sets of keywords. Performing significantly better than all the other methods for all the measures, it improves the performance of the PLSA-MIXED and LSA methods for both normalized score and annotation accuracy measures. The relative annotation accuracy improvement for the Blobs features is 108% with respect to LSA and 32% with respect to PLSA-MIXED (respectively 66% and 33% for the RGB case).

| Method | BLOBS | RGB |
|---|---|---|
| EMPIRICAL | 0.427 (0.016) [36.2] | 0.427 (0.016) [36.2] |
| LSA | 0.521 (0.013) [40.6] | 0.540 (0.011) [37.9] |
| PLSA-SPLIT | 0.273 (0.020) [43.8] | 0.298 (0.022) [36.3] |
| PLSA-MIXED | 0.463 (0.018) [37.2] | 0.473 (0.020) [36.4] |
| PLSA-WORDS | 0.570 (0.013) [31.2] | 0.571 (0.013) [31.3] |

**Table 2: Average maximum normalized score value over the 10 subsets. The variance is given in parantheses and the corresponding average number of keywords predicted is in brackets.**

# 6. CONCLUSION

We proposed a new PLSA-based image auto-annotation system, which uses two linked PLSA models to represent the textual and visual modalites of an annotated image. This allows a different processing of each modality while learning the parameters and makes a truely semantic latent space definition possible. We compared this method to previously proposed systems using different performance measures and showed that this new latent space modeling significantly improves the previous latent space methods based on a concatenated textual+visual representation.

## Acknowledgments

This research has been carried out in the framework of the Swiss NCCR project (IM)2.

# 7. REFERENCES

[1] L. H. Armitage and P. G. Enser. Analysis of user need in image archives. *Journal of Information Science*, 23(4):287–299, 1997.

[2] D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (ACM SIGIR)*, Aug 2003.

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

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

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

[6] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, Nov 2003.

[3] Prec./Recall measures at www.idiap.ch/~monay/acmm04/