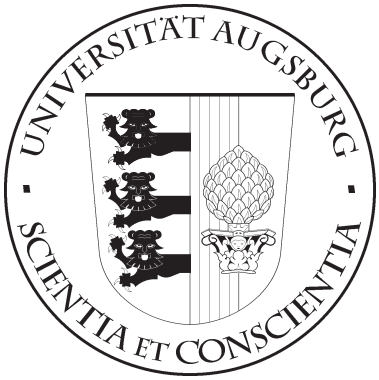
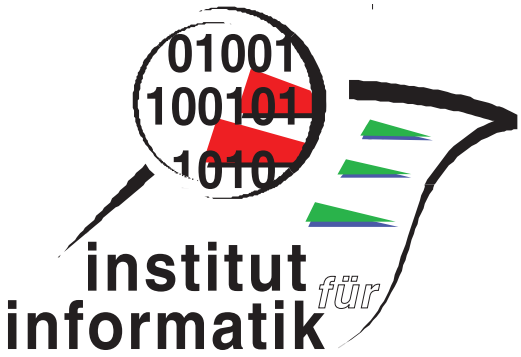
Universität Augsburg



Multilayer pLSA for Multimodal Image Retrieval

R. Lienhart, S. Romberg, E. Hörster

Report 2009-02 April 2009



Institut für Informatik
D-86135 Augsburg

Multilayer pLSA for Multimodal Image Retrieval*

Rainer Lienhart
Multimedia Computing Lab
University of Augsburg
Augsburg, Germany
lienhart@informatik.uni-augsburg.de

Stefan Romberg
Multimedia Computing Lab
University of Augsburg
Augsburg, Germany
romberg@informatik.uni-augsburg.de

Eva Hörster
Multimedia Computing Lab
University of Augsburg
Augsburg, Germany
hoerster@informatik.uni-augsburg.de

ABSTRACT

It is current state of knowledge that our neocortex consists of six layers [10]. We take this knowledge from neuroscience as an inspiration to extend the standard single-layer *probabilistic Latent Semantic Analysis* (*pLSA*) [13] to multiple layers. As multiple layers should naturally handle multiple modalities and a hierarchy of abstractions, we denote this new approach *multilayer multimodal probabilistic Latent Semantic Analysis* (*mm-pLSA*). We derive the training and inference rules for the smallest possible non-degenerated mm-pLSA model: a model with two leaf-pLSAs (here from two different data modalities: image tags and visual image features) and a single top-level pLSA node merging the two leaf-pLSAs. From this derivation it is obvious how to extend the learning and inference rules to more modalities and more layers. We also propose a fast and strictly stepwise forward procedure to initialize bottom-up the mm-pLSA model, which in turn can then be post-optimized by the general mm-pLSA learning algorithm. We evaluate the proposed approach experimentally in a query-by-example retrieval task using 50-dimensional topic vectors as image models. We compare various variants of our mm-pLSA system to systems relying solely on visual features or tag features and analyze possible pitfalls of the mm-pLSA training. It is shown that the best variant of the the proposed mm-pLSA system outperforms the unimodal systems by approximately 19% in our query-by-example task.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*; I.4.8 [**Scene Analysis**]: Sensor fusion; I.4.10 [**Image Representation**]: Statistical

General Terms

Probabilistic image models, hierarchical pLSA, image retrieval, pLSA, SIFT, image tags

Keywords

Image retrieval, multimodal pLSA, hierarchical pLSA, SIFT, tags

*Funded by Deutsche Forschungsgemeinschaft (DFG) under contract number LI 1816/1-1

CIVR '09, July 8-10, 2009 Santorini, GR

1. INTRODUCTION

Many content-based image retrieval systems either solely rely on visual features or on text features to derive a representation of the image content. This is especially true for systems using topic models based on *probabilistic Latent Semantic Analysis* (*pLSA*) [6, 18, 13]. There are good reasons why pLSA is applied to unimodal data: The apparently straightforward application of pLSA to multimodal data by subsuming all words of the various modes (which are generally derived from appropriate features of the respective modality) into one large word set (called vocabulary) frequently does not lead to the expected improvement in retrieval performance. Even mixing words derived from different kinds of features within one domain such as different kinds of visual salient point descriptors (e.g., SIFT [19], SURF [11], Geometric blur [2], or self-similarity feature [22]) using different sampling strategies (e.g., dense versus sparse sampling) does not work satisfactorily with this straightforward application of pLSA.

Thus, we propose a multilayer multimodal pLSA model that can handle different modalities as well as different features within a mode effectively and efficiently. For this we introduce not just a single layer of topics or aspects, but a hierarchy of topics. We explain the overall approach by using the smallest possible non-degenerated mm-pLSA model: a model with two separate sets of topics for data from two different modes and a set of top-level topics that merges the knowledge of the two leaf-topic sets. This resembles somewhat two leaf-pLSAs from two different data modalities that in turn are merged by a single top-level pLSA node, and lends the proposed approach its name: mm-pLSA. From this derivation it is obvious how to extend the learning and inference rules to more modalities and more layers. We also propose a fast and strictly stepwise forward procedure to initialize bottom-up the mm-pLSA model that leads to much better learning results of the mm-pLSA learning algorithm compared to random initialization.

The paper is organized as follows. In Section 2 we first describe the model of the standard pLSA algorithm (Sub-

section 2.1) as well as how to learn a pLSA model in general (Subsection 2.2) and specifically from the visual features (Subsection 2.3) and tag features (Subsection 2.4). Classification of a new image or text document is also addressed. Then, Section 3 presents the core novelty of our work in detail: the *multilayer multimodal probabilistic Latent Semantic Analysis* model (*mm-pLSA*). It starts in Subsection 3.1 with a motivation and a detailed explanation of the model, before we derive the training and inference steps in Subsection 3.2. A heuristic for fast and good initialization of the multilayer multimodal pLSA model is presented in Subsection 3.3 and carefully evaluated in Section 4 on a large scale database consisting of 246,347 images downloaded from Flickr. Our proposed mm-pLSA based image retrieval system is compared to systems relying solely on visual features [18] or tag features as well as to a pLSA-based system with the combined vocabulary set from the visual and tag domain. Section 5 details related work, before Section 6 concludes the paper.

2. STANDARD PLSA

2.1 Motivation and Model

The pLSA was originally devised by T. Hofmann in the context of text document retrieval, where words constitute the elementary parts of documents [13]. Applied to images, each image represents a single visual document. pLSA can be applied directly to image tags, as image tags consist of words. However, for our visual features we need comparable elementary parts called visual words. For the moment we assume that all features we computed in a given mode are somehow mapped to words in that mode. Details of the mapping from the visual features to the mode-specific words are given in Subsection 2.3. For now we just assume that we have words.

The key concept of the pLSA model is to map the high-dimensional word distribution vector of a document to a lower dimensional *topic vector* (also called *aspect vector*). Therefore pLSA introduces a latent, i.e. unobservable, topic layer between the documents (i.e., images here) and the words. It is assumed that each document consists of a mixture of multiple topics and that the occurrences of words (i.e., visual words in images or tags of images, respectively) is a result of the topic mixture. This generative model is expressed by the following probabilistic model:

$$P(d_i, w_j) = P(d_i) \sum_K P(z_k|d_i)P(w_j|z_k) \qquad (1)$$

where $P(d_i)$ denotes the probability of a document d_i of the database to be picked, $P(z_k|d_i)$ the probability of a topic z_k given the current document, and $P(w_j|z_k)$ the probability of a visual word w_j given a topic. The model is graphically depicted in Fig. 1. N_i denotes the number of words of which each of the M documents consists. It is important not to confuse N_i, the number of words in document d_i, with N, the number of words in the vocabulary.

Once a topic mixture $P(z_k|d_i)$ is derived for each document d_i, a high-level representation based on the respective mode the words belong to has been found. At the same time this representation is of low dimensionality as we commonly choose the number of concepts in our model to be much smaller than the number of words. The K-dimensional topic vector can be used directly in a query-by-example retrieval task, if we measure document similarity by computing the L_1 or cosine distance between topic vectors of different documents.

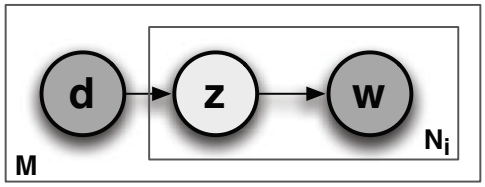


Figure 1: Standard pLSA-Model.

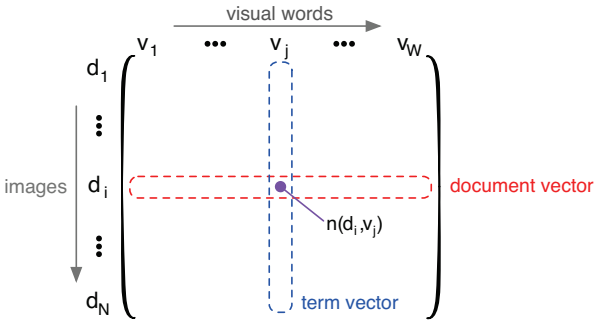


Figure 2: Term-document matrix.

2.2 Training and Inference

Computing the *term-document matrix* of the training corpus is a prerequisite for deriving the pLSA model (see Fig. 2). Each row i in the term-document matrix $[n(d_i, w_j)]_{i,j}$ describes the absolute count with which word w_j (also called a term) occurs in document d_i. The terms are taken from a predefined dictionary consisting of N terms. The number of documents is M. Thus $n(d_i, w_j)$ specifies the number of times the word w_j occurred in document d_i. Note that by normalizing each document vector to 1 using the L1-norm, the document vector $(n(d_i, w_1), ..., n(d_i, w_N))$ of d_i becomes the estimated mass probability distribution $P(w_j|d_i)$.

We learn the unobservable probability distribution $P(z_k|d_i)$ and $P(w_j|z_k)$ from the data using the Expectation-Maximization-Algorithm (EM-Algorithm) [7, 13]:

E-Step:

$$P(z_k|d_i, w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{l=1}^{K} P(w_j|z_l)P(z_l|d_i)} \qquad (2)$$

M-Step:

$$P(w_j|z_k) = \frac{\sum_{i=1}^{M} n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{j=1}^{N}\sum_{i=1}^{M} n(d_i, w_j)P(z_k|d_i, w_j)} \qquad (3)$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{N} n(d_i, w_j)P(z_k|d_i, w_j)}{n(d_i)} \qquad (4)$$

Given a new test image d_{test}, we estimate the topic probabilities from the observed words. The sole difference between inference and learning is that the K learned conditional word distributions $P(w_j|z_k)$ are never updated during inference.

Thus, only eq. (2) and eq. (4) are iteratively updated during inference.

2.3 Visual pLSA-Model

The first step in building a bag-of-words representation for the visual content of images is to extract visual features from each image. In our case we apply dense sampling with a vertical and horizontal step size of 10 pixels across the image pyramid created with a scale factor of 1.2 to extract local image features at regular grid keypoints. SIFT descriptors [19] computed over a region of 41×41 pixels are used to describe the grayscale image region around each keypoint in a scale and orientation invariant fashion. Although we use SIFT features in this work, any other feature could be used instead.

Next the 128-dimensional real-valued local image features have to be quantized into discrete visual words to derive a finite vocabulary. Quantization of the features into visual words is performed by using a vocabulary tree [21] in order to support large vocabulary sizes. The vocabulary tree is computed by repeated k-means clusterings that hierarchically partition the feature space. This hierarchical approach overcomes two major problems of the traditional direct k-means clustering in cases where k is large. Firstly clustering is more efficient during visual word learning and secondly the mapping of visual features to discrete words is way faster than using a plain list of visual words.

Once the visual vocabulary of size N^v is determined we map each feature vector of an image to its closest visual word. Therefore we query the vocabulary tree for each extracted feature, and the best matching visual word ID is returned. This ID in turn is used to compute the document vector that holds the counts of the occurrences of visual words in the corresponding image by incrementing the associated word count (see Fig. 3).

Note that this image content description does not preserve any spatial relationship between the occurrences of the visual words. The co-occurrence vectors of all training images are used to train the pLSA model. Once the pLSA model is learned it can be applied to all images in the database, hence deriving a vector representation for each image, where the vector elements denote the degree to which an image depicts a certain visual topic.

2.4 Tag-based pLSA-Model

The second modality we considered are tags – the free-text annotations provided by the image author. We assume that all of the images in our database have been tagged by their authors. Besides a single word, a tag can also be a phrase or a sentence. However, in this work we treat each word of the image annotations separately. Thus, in the following the term *tag* denotes a single word and is used interchangeably with "word" and "term".

As we use Flickr images to evaluate our multilevel multimodal pLSA model, it is important to note that these tags reflect the photographer/author's personal view with respect to the uploaded image. Thus, in contrast to carefully annotated image databases traditionally used for learning combined image and tag models [1], these image tags from Flickr are in many cases subjective, ambiguous, and do not necessarily describe the image content shown [17, 18]. This makes it difficult to use the tags directly for retrieval purposes and thus some preprocessing is required.

breakfast	eat food meal
house	construction home object structure
love	emotion state

Table 1: Examples for hypernyms (right) found in Wordnet for the words left.

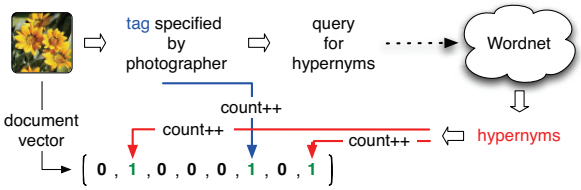


Figure 4: Building a document vector from tags and their hypernyms by assuming that both tag and hypernyms are present in the vocabulary.

To apply a pLSA model to tags we need to define a finite vocabulary first. Building the vocabulary starts with listing all tags that have been used more than N_1^{thresh} times and by at least N_2^{thresh} different users. This heuristics enforces that all rarely used tags are neglected. Note that a tag is also rarely used, if only a few users have used it independent of the actually count. We further filter the list by discarding all tags that contain numbers and by splitting tags at underscores into separate words. In a last filtering step all words within the vocabulary are checked whether they are known by Wordnet [8]. Wordnet is a lexical database of English. Only words that exist according to Wordnet are assimilated into the final vocabulary.

Once the tag vocabulary of size N^t is defined, a co-occurrence table is built by counting the tag occurrences for each image. However, before forming the document vector we expand the list of tags associated with an image by using Wordnet to enrich the annotations. For each image we query Wordnet for the semantic parents of the tags specified by the author. This is done to emphasize semantic features rather than just the simple word count features.

Semantic parents for a word can easily be extracted from Wordnet, as each word in Wordnet is associated with hypernyms[1]. A word's hypernyms denote its parents that express the specific concept of the tag more generally. Table 1 shows hypernyms (right) for some example tags (left). As these hypernyms build a hierarchy and form a tree structure, we add the hypernyms up to three levels above in the hierarchy of the tag itself into the tag list of the corresponding image. Thus, while counting tag occurrences for each image in order to build the document vector, these parents are included in our model by counting them as if they were present in the list of tags. This procedure is visualized in Fig. 4. In case the vocabulary does not contain a tag or hypernyms used for annotation, the word is simply ignored.

In our experiments we set the parameters for the tag vocabulary to $N_1^{thresh} = 18$ and $N_2^{thresh} = 10$ resulting in a vocabulary size of 2421 words. Most images in our database have between 5 and 15 tags plus hypernyms as can be derived from Fig. 5. The number of tags for some images is however unreasonably large as users labeled images with whole sentences or phrases.

[1] Y is a hypernym of X if every X is a (kind of) Y.

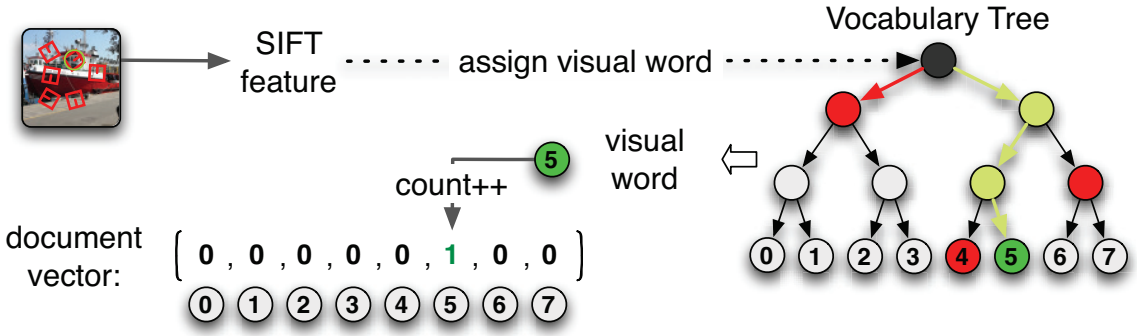


Figure 3: Quantization of features into discrete visual words.

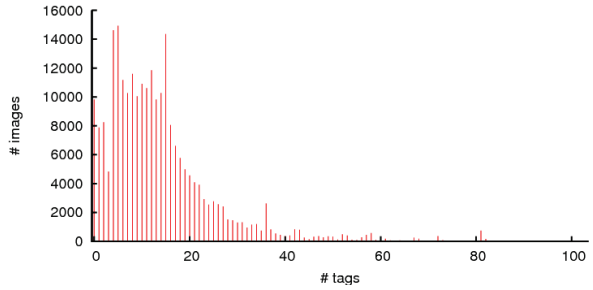


Figure 5: Histogram of the total number of tags plus hypernyms for each image within our database.

3. MULTILAYER MULTIMODAL PLSA

3.1 Motivation and Model

In recent years pLSA has been applied successfully to unimodal data such as text [13], image tags [20], or visual words [16]. However, combining two modes such as visual words and image tags is challenging. The obvious approach of simply concatenating the two associated term-document matrices $N_{M\times N^v}$ and $N_{M\times N^t}$ into $N_{M\times(N^v+N^t)}$ and then applying standard pLSA usually does not lead to the desired retrieval improvements. One reason is the difference in the order of magnitude with which words occur in the respective mode. For instance, a few thousand to ten thousand features per image are usually computed from images that are resized to having roughly the same number of dense samples while preserving the image's aspect ratio. In contrast, most images are annotated with fewer than 20 tags. Compensating between the differences in the order of the magnitude by some kind of normalization is possible, but will require a lot of testing to determine an appropriate weighting factor between the different modes since the actual importance of each mode must also be taken into account. Another reason may be the difference in the size of the respective vocabularies. In contrast, a well-founded mathematical approach with top-level topics will solve this issue effectively and efficiently. Some empirical evidence for these claims will be given in section 4.

Our basic idea is to apply pLSA in a first step to each mode separately, and in a second step concatenate the derived topic vectors of each mode to learn another pLSA on top of that (see Fig. 8). While we describe this layering of multiple pLSAs only for two leaf-pLSAs and a node pLSA, it is obvious that the proposed pLSA layering can be extended to more than two layers and applied to more than just two leaf-pLSAs.

The smallest possible multilayer multimodal pLSA model (mm-pLSA) considering two modes with their respective observable word occurrences and hidden topics as well as a single top-level of hidden aspects is graphically depict in Fig. 6. Every word of mode x (here: $x \in \{v, t\}$ with v standing for *visual* and t for *text*) occurring in document d_i is generated by an unobservable model document:

- Pick a document d_i with prior probability $P(d_i)$
- For each visual word in the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top}|d_i)$
  - Select a visual topic z_k^v with probability $P(z_k^v|z_l^{top})$
  - Generate a visual word w_m^v with probability $P(w_m^v|z_k^v)$
- For each tag associated with the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top}|d_i)$
  - Select an image tag topic z_p^t with probability $P(z_p^t|z_l^{top})$
  - Generate an image tag w_n^t with probability $P(w_n^t|z_p^t)$

Thus the probability of observing a visual word w_m^v or a tag w_n^t in document d_i is

$$P(d_i, w_m^v) = \sum_{l=1}^{L}\sum_{k=1}^{K} P(d_i)P(z_l^{top}|d_i)P(z_k^v|z_l^{top})P(w_m^v|z_k^v) \quad (5)$$

$$P(d_i, w_n^t) = \sum_{l=1}^{L}\sum_{p=1}^{P} P(d_i)P(z_l^{top}|d_i)P(z_p^t|z_l^{top})P(w_n^t|z_p^t). \quad (6)$$

An important aspect of this model is that every image consists of one or more part aspects in each mode, which in turn are combined to one or more higher-level aspects. This is very natural since images consist of multiple objects parts and multiple objects. The multilayer multimodal pLSA can model this fact effectively – much better than a single layer pLSA. Furthermore this model is in better correspondence with current belief in hierarchical recurrent cortex models of our brain [10].

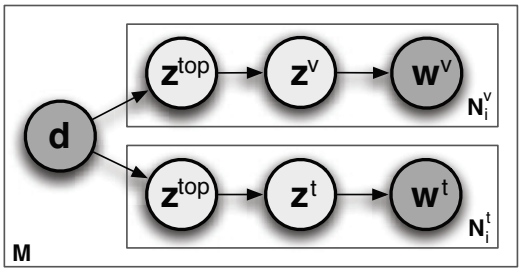


Figure 6: The new multi-layer multi-model pLSA model illustrated by combining two modalities.

3.2 Training and Inference

Given our word generation model (see Fig. 6) with its implicit independence assumption between generated words, the likelihood L of observing our database consisting of the observed pairs (d_i, w_m^v) and (d_i, w_n^t) from both modes is given by

$$L = \prod_{i=1}^{M} \left[\prod_{m=1}^{N^v} P(d_i, w_m^v)^{n(d_i, w_m^v)} \prod_{n=1}^{N^t} P(d_i, w_n^t)^{n(d_i, w_n^t)} \right]. \quad (7)$$

Taking the log to determine the log-likelihood l of the database

$$l = \sum_{i=1}^{M} \left[\sum_{m=1}^{N^v} n(d_i, w_m^v) \log P(d_i, w_m^v) + \sum_{n=1}^{N^t} n(d_i, w_n^t) \log P(d_i, w_n^t) \right] \quad (8)$$

and plugging eq. (5) and (6) in to eq. (8), it becomes apparent that there is a double sum inside of both *log*s making direct maximization with respect to the unknown probability distributions difficult. Therefore, we learn the unobservable probabilities distribution $P(z_l^{top}|d_i)$, $P(z_k^v|z_l^{top})$, $P(z_p^t|z_l^{top})$, $P(w_m^v|z_k^v)$ and $P(w_n^t|z_p^t)$ from the data using the Expectation-Maximization-Algorithm (EM-Algorithm) [7]. Introducing the indicator variables

$$\triangle c_{lk} = \begin{cases} 1 & \text{if the pair } (d_i, w_m^v) \text{ was generated by } z_l^{top} \text{ and } z_k^v \\ 0 & \text{otherwise} \end{cases}$$

$$\triangle d_{lp} = \begin{cases} 1 & \text{if the pair } (d_i, w_p^t) \text{ was generated by } z_l^{top} \text{ and } z_p^t \\ 0 & \text{otherwise} \end{cases}$$

the complete data likelihood L_c, that is the data likelihood assuming that d_i, w_n^z, w_m^v, $\triangle c_{lk}$, and $\triangle d_{lp}$ are observable, is given by

$$L_c = \prod_{i=1}^{M} [\prod_{m=1}^{N^v} P(d_i, w_m^v, \triangle c)^{n(d_i, w_m^v)} \prod_{n=1}^{N^t} P(d_i, w_n^t, \triangle d)^{n(d_i, w_n^t)}]$$

with

$$\triangle c = (\triangle c_{11}, ..., \triangle c_{1K}, ..., \triangle c_{LK}) \quad (9)$$

$$\triangle d = (\triangle d_{11}, ..., \triangle d_{1K}, ..., \triangle d_{LP}) \quad (10)$$

$$P(d_i, w_m^v, \triangle c) = \prod_{l=1}^{L} \prod_{k=1}^{K} P(d_i) P(z_l^{top}|d_i) P(z_k^v|z_l^{top}) P(w_m^v|z_k^v)^{\triangle c_{lk}} \quad (11)$$

$$P(d_i, w_n^t, \triangle d) = \prod_{l=1}^{L} \prod_{p=1}^{P} P(d_i) P(z_l^{top}|d_i) P(z_p^t|z_l^{top}) P(w_n^t|z_p^t)^{\triangle d_{lp}} \quad (12)$$

Unlike in eq. (8), we now only have product terms in the complete likelihood L_c, thus its log-likelihood can easily be termined and maximized[2], resulting in the following expectation (E-step) and maximization (M-step) solution:

E-Step:

We estimate the unknown indicator variables $\triangle c_{lk}$ conditioned on the observable variables d_i and w_m^v by computing their expected value:

$$\begin{aligned} c_{lk}^{im} &:= E(\triangle c_{lk} | d_i, w_m^v) \\ &= P(\triangle c_{lk} = 1 | d_i, w_m^v) \cdot 1 + P(\triangle c_{lk} = 0 | d_i, w_m^v) \cdot 0 \\ &= P(\triangle c_{lk} = 1 | d_i, w_m^v) \cdot 1 \\ &= \frac{P(d_i, w_m^v, \triangle c_{lk} = 1)}{P(d_i, w_m^v)} \\ &= \frac{P(d_i) P(z_l^{top}|d_i) P(z_k^v|z_l^{top}) P(w_m^v|z_k^v)}{\sum_{l=1}^{L} \sum_{k=1}^{K} P(d_i) P(z_l^{top}|d_i) P(z_k^v|z_l^{top}) P(w_m^v|z_k^v)}. \end{aligned} \quad (13)$$

Analogously we estimate the unknown indicator variables $\triangle d_{lp}$ conditioned on the observable variables d_i and w_n^t by computing their expected value:

$$\begin{aligned} d_{lp}^{in} &:= E(\triangle d_{lp} | d_i, w_n^t) \\ &= \frac{P(d_i) P(z_l^{top}|d_i) P(z_p^t|z_l^{top}) P(w_n^t|z_p^t)}{\sum_{l=1}^{L} \sum_{k=1}^{K} P(d_i) P(z_l^{top}|d_i) P(z_p^t|z_l^{top}) P(w_n^t|z_p^t)} \end{aligned} \quad (14)$$

M-Step:

For legibility of the M-step estimates, we set

$$\gamma_{lk}^{im} := n(d_i, w_m^v) c_{lk}^{im} \quad (15)$$

$$\delta_{lp}^{in} := n(d_i, w_n^t) d_{lp}^{in} \quad (16)$$

which is the expected probability of observing a pair (d_i, w_m^v) multiplied with the actual number of occurrences and get:

$$P(d_i)^{new} = \frac{\sum_{m=1}^{N^v} n(d_i, w_m^v) + \sum_{n=1}^{N^t} n(d_i, w_n^t)}{\sum_{i=1}^{M} \left(\sum_{m=1}^{N^v} n(d_i, w_m^v) + \sum_{n=1}^{N^t} n(d_i, w_n^t) \right)} \quad (17)$$

$$P(z_l^{top}|d_i)^{new} = \frac{\sum_{m=1}^{N^v} \sum_{k=1}^{K} \gamma_{lk}^{im} + \sum_{n=1}^{N^t} \sum_{p=1}^{P} \delta_{lp}^{in}}{\sum_{l=1}^{L} \left(\sum_{m=1}^{N^v} \sum_{k=1}^{K} \gamma_{lk}^{im} + \sum_{n=1}^{N^t} \sum_{p=1}^{P} \delta_{lp}^{in} \right)} \quad (18)$$

$$P(z_k^v|z_l^{top})^{new} = \frac{\sum_{i=1}^{M} \sum_{m=1}^{N^v} \gamma_{lk}^{im}}{\sum_{k=1}^{K} \sum_{i=1}^{M} \sum_{m=1}^{N^v} \gamma_{lk}^{im} + \sum_{p=1}^{P} \sum_{i=1}^{M} \sum_{n=1}^{N^t} \delta_{lp}^{in}} \quad (19)$$

$$P(z_p^t|z_l^{top})^{new} = \frac{\sum_{i=1}^{M} \sum_{n=1}^{N^t} \delta_{lp}^{in}}{\sum_{k=1}^{K} \sum_{i=1}^{M} \sum_{m=1}^{N^v} \gamma_{lk}^{im} + \sum_{p=1}^{P} \sum_{i=1}^{M} \sum_{n=1}^{N^t} \delta_{lp}^{in}} \quad (20)$$

[2]A complete derivation of the EM-update equation for this multilayer multimodel pLSA model can be found at http://mmc36.informatik.uni-augsburg.de/mediawiki-1.11.2/images/7/7f/CIVR2009-EM-derivation.pdf

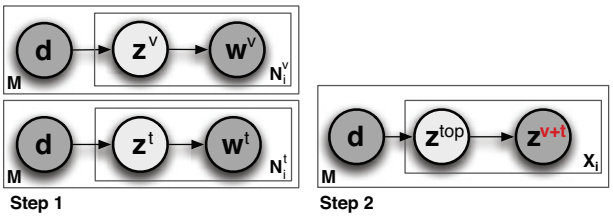


Figure 7: The fast initialization of the multimodal pLSA model computed in two separate steps.

$$P(w_m^v|z_k^v)^{new} = \frac{\sum_{i=1}^{M}\sum_{l=1}^{L}\gamma_{lk}^{im}}{\sum_{m=1}^{N^v}\sum_{i=1}^{M}\sum_{l=1}^{L}\gamma_{lk}^{im}} \quad (21)$$

$$P(w_n^t|z_p^t)^{new} = \frac{\sum_{i=1}^{M}\sum_{l=1}^{L}\delta_{lp}^{in}}{\sum_{n=1}^{N^t}\sum_{i=1}^{M}\sum_{l=1}^{L}\delta_{lp}^{in}} \quad (22)$$

Clearly eq.(17) is constant across all iterations and must not be recomputed.

Given a new test image d_{test}, we estimate the top-level aspect probabilities $P(z_l^{top}|d_{test})$ with the same E-step equations as for learning and eq. (18) for $P(z_l^{top}|d_{test})$ as the M-step. The probabilities of $P(z_k^v|z_l^{top})$, $P(z_p^t|z_l^{top})$, $P(w_m^v|z_k^v)$ and $P(w_n^t|z_p^t)$ have been learned from the corpus and are kept constant during inference.

3.3 Fast Initialization

More complicated probabilistic models always come with an explosion in required training time. This issue is becoming more severe, the more layers and the more pLSAs are aggregated into higher-level pLSAs. Thus, we suggest to compute a decent initial estimation of the conditional probabilities in a strictly stepwise forward procedure (see also Figure 8). For the smallest two-leaf high-level aspect model this procedure first computes an independent pLSA for each mode on the lowest level. The aspects are only linked through the documents, ie., the same images (see Step 1 in Fig. 7). Next the computed aspect of all modes are taken as the observed words at the next higher level (see Step 2 in Fig. 7). This procedure can continue until the top-level aspect vector is learned. The final representation, the top-level aspect distribution for each document, describes each image as a "distribution over topic distributions" and thereby fuses the visual pLSA model and the tag pLSA model. As we will show in the experimental results, this fast initialisation already produces a decent model. It can be further be improved by appying the EM-algorithm as stated in section 3.2 to the complete model after initialising it with the strictly forward computed solution. This will futher improve the solution. An overview of an image retrieval system based on this idea is shown in Figure 8.

4. EXPERIMENTAL EVALUATION

4.1 Setup

All presented systems were experimentally evaluated on a dataset consisting of 246,347 geotagged Flickr images associated with at least a single tag and belonging to one of 12 different categories. The database had not been cleaned or post-processed [18].

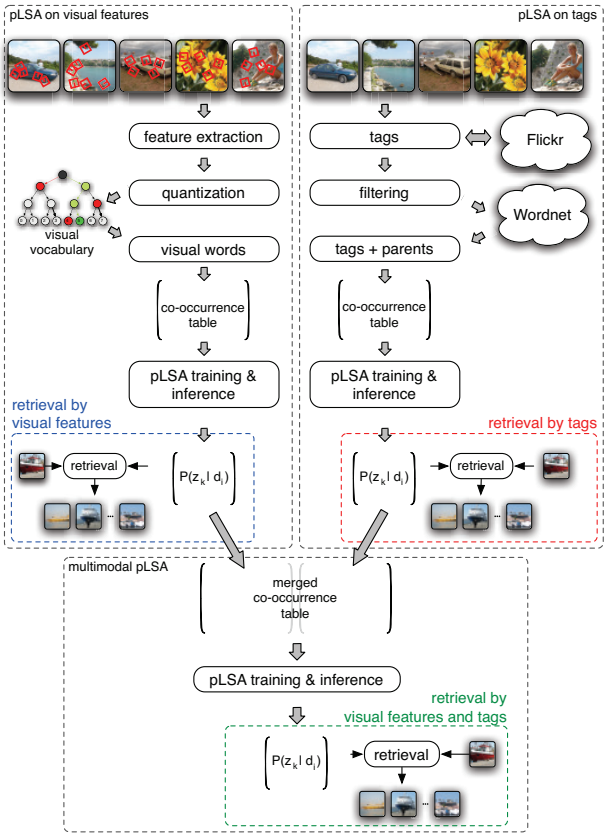


Figure 8: Overview of the retrieval system based on our fast initialization strategy.

A visual vocabulary of size 10,000 was computed. We learned two 50-topic pLSA model: one for tag features and one for visual features. Each pLSA model, independent of whether a conventional unimodal or a multilevel multimodal pLSA model was trained with 5000 images except the tag model that was learned from 10,000 images. The training corpus for tags had been widened to cover more of the "tag space" as the vectors that count tag occurrences were very sparse (see Fig. 5).

The fast initialization of the mm-pLSA mapped the two 50-dimensional image representations computed by the two base models (based on visual features and tags) to a multimodal topic distribution over 50 "super" topics. The randomly initialized mm-pLSA and its optimized version with the general mm-pLSA learning algorithm directly computed a model with 50 topics. The number of iterations used during training and inference varied. All models were computed using 450 iterations, except the mm-pLSA with the fast initialization method. This model was computed using 50 iterations, since we already had a good starting point.

The only probability distribution computed during inference was the probability distribution $P(z_l^{top}|d_i)$ of the top-level topics given the document. Therefore the EM-algorithm converged faster than during training and the

number of iterations was reduced. For the inference of these topic distributions we used 200 iterations with the visual-based pLSA, the tag-based pLSA, the concatenated topic-based pLSA, the fast initialization of the mm-pLSA, and the mm-pLSA with random initialization. 50 iterations were used for the mm-pLSA with fast initialization.

We evaluated all the systems in a query-by-example task and evaluated the results by a user study with 8 users. 60 query images were selected and the L_1 distance was used to find their most similar images. The users were asked to rate the 19 closest results to each of our query images. Note that we always showed the images without their associated tags as we evaluated a query-by-image-example system. We used the following scoring to get a quantitative performance measure: An image considered being similar received 1 point, an image considered somewhat similar received 0.5 points. All other images got 0 points. A mean score was calculated for each user; the mean over all eight users' means yielded the final score of the system being evaluated.

4.2 Results

The results of our experiment are shown in Figure 9. The first two experiments denote the performance of the systems based solely on visual features or tags. "Concatenated pLSA" denotes the pLSA model computed from merging the words from the visual domain as well as the tag domain into one vocabulary. The straight-forward approach of applying a third pLSA model on top of the two base models is shown as "mm-pLSA (fast init only)", while the mm-pLSA that is initialized randomly or with the outcome of the fast initialization is denoted as "mm-pLSA (random init)" or "mm-pLSA (fast init)", respectively.

It can be seen that the systems relying solely on tags perform better than the system relying solely on visual features. In contrast, the system aiming to fuse the modalities by concatenating the occurrence counts performs worse than the system based on tags only. Expectedly its performance is the same as of the system based on visual features only, as the few items added to the co-occurrence matrix by concatenating the tag occurrences to the visual feature occurrences are unlikely to have a major impact on the learned pLSA model training. Both mm-pLSA models with fast initialization only and with optimizing the already good initialization outperformed these systems by up to 24% which confirms the expected superior performance of multimodal multilevel models.

The randomly mm-pLSA performed better "mm-pLSA (fast init only)", but worse than the "mm-pLSA (with fast init)". This is in line with our expectations: we expected the random initialized model to perform inferior to its well initialized counterpart. It should be noted that as the EM-algorithm already starts from a relatively good solution, the number of training iterations can be small. Therefore the mm-pLSA with fast initialization is quite fast and effective.

5. RELATED WORK

Topic models have been used in several previous works in order to derive a low dimensional image description suitable for large-scale image retrieval, for example [18] used probabilistic Latent Semantic Analysis (pLSA) [13] based models, [15] applied Latent Dirichlet Allocation (LDA) [5] to derive a topic representation and [9] adopted the Correlated Topic Model (CTM) [3]. However all of the previous

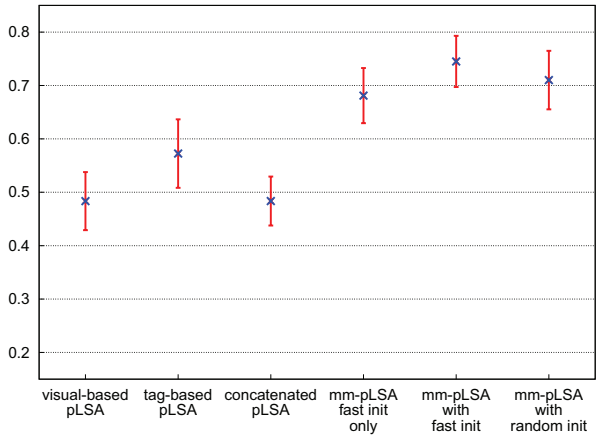


Figure 9: Scores for our different retrieval systems. Vertical bars mark the standard deviation between the users' means.

mentioned works built their image representation solely on visual features.

In [1],[4] and [20] the authors propose topic models to model annotated image databases. They use the models to automatically annotate images and/or image regions. One big difference of our work to those previous works is that we aim to build an image retrieval system instead of annotating images. Moreover the image database we use for learning and retrieval is a real world, large scale database in contrast to the almost noise free database (COREL database) that was used in the above works for learning and testing. Thus in our case the tags associated with an image do not necessarily refer to the visual content shown. For example, they may also denote the time, date, place or circumstances where the picture was taken. This makes models which try to associate image regions directly with tags difficult to learn and apply.

Our approach uses an hierarchical model as we have more than one topic layer. In [23] the authors adapt the Hierarchical Latent Dirichlet Allocation (hLDA) model, which has been developed originally for the unsupervised discovery of topic hierarchies in text, to the visual domain. They use the model for object classification and segmentation. However their model only accounts for one modality, visual features. Moreover appropriate initialization of the complex model is difficult. Another example for a hierarchical model for image content are deep networks [12][14] with which – on a very high-level point of view – we share the stepwise forward initialization and subsequent optimization.

6. CONCLUSION

A very general scheme for multilayer multimodal probabilistic Latent Semantic Analysis has been proposed. It naturally extends the single-layer pLSA to the concept of layered or hierarchical topics – something we humans have to deal with on a daily basis. It also allows grasping concepts across different modalities. The proposed fast initialization technique makes the mm-pLSA very practical and computable. The overall approach was evaluated in a query-by-example image retrieval scenario by users and outper-

formed unimodal pLSA by more than 19% percent. It is apparent, that the simple two leaves, one node instance of that model was just an example and can be extended to complete tree structures. Thus the mm-pLSA shows huge promise for future research.

Although the results of the randomly initialized mm-pLSA are disappointing, the results point to an important issue of layered or hierarchical models in general. They are highly effected by good initializations, which our heuristic provides. Without it the local maximization procedure gets trapped too easily. Future work will focus on how to integrate a relevancy measure for each internal topic set in the model such that a better learning between different modes can be achieved.

7. REFERENCES

[1] K. Barnard, P. Duygulu, D. Forsyth, D. M. Blei, T. Hofmann, T. Poggio, and J. Shawe-taylor. Matching words and pictures. *Journal of Machine Learning Research 3 (2003) 1107–1135*, 2003.

[2] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, pages 26–33, Washington, DC, USA, 2005. IEEE Computer Society.

[3] D. Blei and J. Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems 18*, pages 147–154. 2006.

[4] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, 2003.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[6] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification via pLSA. *Proceedings of the European Conference on Computer Vision*, 3954:517–530, 2006.

[7] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[8] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[9] T. Greif, E. Hörster, and R. Lienhart. Correlated topic models for image retrieval. In *Technical Report TR2008-09, University of Augsburg*, 2008.

[10] J. Hawkins and S. Blakeslee. *On Intelligence*. Times Books, 2004.

[11] T. T. Herbert Bay and L. V. Gool. Surf: Speeded up robust features. *Proceedings of the 9th European Conference on Computer Vision, Springer LNCS*, 3951(1):404–417, 2006.

[12] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[13] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42, Numbers 1-2:177–196, 2001.

[14] E. Hörster and R. Lienhart. Deep networks for image retrieval on large-scale databases. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, pages 643–646, New York, NY, USA, 2008. ACM.

[15] E. Hörster, R. Lienhart, and M. Slaney. Image retrieval on large-scale image databases. In *Proc. ACM CIVR'07*, pages 17–24, 2007.

[16] E. Hörster, R. Lienhart, and M. Slaney. Continuous visual vocabulary modelsfor plsa-based scene recognition. In *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 319–328, New York, NY, USA, 2008. ACM.

[17] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 631–640, New York, NY, USA, 2007. ACM.

[18] R. Lienhart and M. Slaney. pLSA on large scale image databases. *IEEE International Conference on Acoustics, Speech and Signal Processing 2007 (ICASSP 2007)*, Vol. IV:1217–1220, 2007.

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60 (2):91–110, 2004.

[20] F. Monay and D. Gatica-Perez. Plsa-based image auto-annotation: constraining the latent space. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 348–351, New York, NY, USA, 2004. ACM.

[21] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:2161–2168, 2006.

[22] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition 2007 (CVPR'07)*, June 2007.

[23] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros. Unsupervised discovery of visual object class hierarchies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.