Universität Augsburg

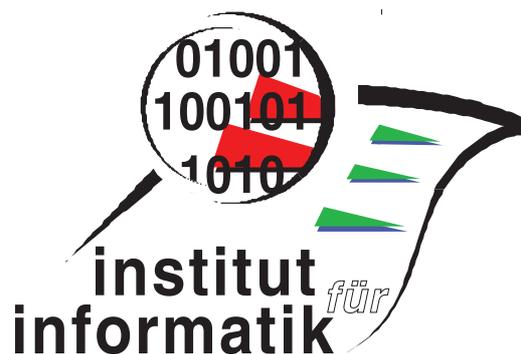# Multimodal pLSA on Visual Features and Tags

S. Romberg, E. Hörster, R. Lienhart

Institut für Informatik
D-86135 Augsburg

MULTIMODAL PLSA ON VISUAL FEATURES AND TAGS

Stefan Romberg, Eva Hörster, Rainer Lienhart

Multimedia Computing Lab
University of Augsburg
Augsburg, Germany

ABSTRACT

This work studies a new approach for image retrieval on large-scale community databases. Our proposed system explores two different modalities: visual features and community-generated metadata, such as tags. We use topic models to derive a high-level representation appropriate for retrieval for each of our images in the database. We evaluate the proposed approach experimentally in a query-by-example retrieval task and compare our results to systems relying solely on visual features or tag features. It is shown that the proposed multimodal system outperforms the unimodal systems by approximately 36%.

Index Terms— image retrieval, multimodal pLSA, SIFT, tags

1. INTRODUCTION

Many content-based image retrieval systems solely rely on visual features to derive a representation of the image content. However, nowadays images are often stored in and retrieved from large-scale community databases such as Flickr. In many of those community databases the images are associated with different kinds of metadata, e.g. camera data (such as focal length), image title or author. This additional information can be used to improve the performance of visual feature-based image retrieval.

In this work we explore one specific type of metadata, tags specified by the photographer/author of the image. Community databases allow authors to use tags to label their images with keywords in order to describe them. These tags usually reflect the users 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], tags associated with images in such community databases are in many cases ambiguous and do not necessarily describe the image content shown. This makes it difficult to use the tags directly for retrieval purposes and more sophisticated models need to be developed.

Our approach uses probabilistic Latent Semantic Analysis (pLSA) models [2] to build a high level representation appropriate for retrieval which considers images as mixtures of topics. Similar to [3] we build a pLSA model based on visual features. However we extend the approach of [3] and do not use this representation directly for retrieval. Instead we compute a second pLSA model based on tag features. Finally we join both modalities by learning a third pLSA model on the already derived topic mixtures and thus derive a multimodal high-level image representation appropriate for retrieval.

We evaluate our approach on a large scale database consisting of 246,347 images downloaded from Flickr by comparing the proposed multimodal system to systems relying solely on visual features [3] or tag features.

The paper is organized as follows. In the next two sections we will first describe how a pLSA model is learned from visual (Section 2) and tag features (Section 3). Section 4 presents our multimodal retrieval system and in Section 5 we evaluate our proposed approach.

2. VISUAL FEATURES

Learning a pLSA model from visual features starts with representing each image as a collection of visual words from a discrete and finite visual vocabulary, the so called bag of word model. The occurrences of visual words in an image are hereby counted into a co-occurrence vector, also called document vector. Note that this image content description does not preserve any spatial relationship between the occurrence of the visual words. The co-occurrence vectors of all images then build the co-occurrence table which is used to train the pLSA model. Once the pLSA model is learned it can be applied to all images in the database thus deriving a vector representation for each image, where the vector elements denote the degree to which an image depicts a certain topic.

2.1. Building the Co-occurrence Table

The first step while building a bag-of words representation for our images is to extract visual features from each image. In our case we extract local image features at keypoints found at extrema of the Difference of Gaussian pyramid. Scale Invariant Feature Transform (SIFT) descriptors [4] are used to describe the grayscale image region around each keypoint in a scale and orientation invariant fashion. Although we use
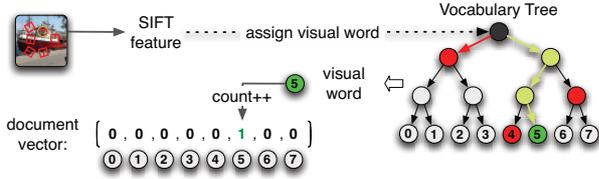
Fig. 1. Quantization of features into discrete visual words.

SIFT features in this work here, any other features could be used in our model 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 [5]. 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. First clustering is more efficient during visual word learning and second the mapping of visual features to discrete words is way faster than using a plain list of visual words.

Once the visual vocabulary is determined we map each feature vector 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 is then used to access the document vector that holds the count of the occurrences of visual words in the corresponding image and the according word count is incremented (Figure 1).

2.2. pLSA - Training & Inference

Having computed a document vector for each image we use the pLSA [2] model to derive high level visual features. The key concept of the pLSA model is thereby to map the high-dimensional discrete count vectors to lower dimensional topic vectors. Therefore pLSA introduces a latent, i.e. unobservable, topic layer between images and the visual words. It is assumed that each image consists of a mixture of multiple topics and that the occurrences of visual words in images is a result of the topic mixture. This results in 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 of the database to be picked, $P(z_k|d_i)$ is the probability of a topic given the current document and the probability of a visual word given a topic is denoted by $P(w_j|z_k)$.

Although latent topics describe the content of images, only the occurrence of visual words itself in images can be observed in practice. To learn the pLSA model the Expectation-Maximization algorithm [6][2] is applied. Note that the model learning is completely unsupervised and therefore the topics itself are defined unsupervised as well. As we usually train our model only on a subset of the entire database we need to be able to map a new image to the model. This is done by the so called fold-in technique [2].

Once a topic mixture $P(z|d)$ is derived for each image, a high-level representation based on the visual features has been found. At the same time it is a dimensionality reduction as we commonly choose the number of concepts in our model smaller than the number of visual words.

The K-dimensional topic vector can be used directly for a query-by-example retrieval task if we measure image similarity by computing the L_1 or Cosine distance between topic vectors of different images. However the system proposed here aims to fuse different kinds of modalities in order to improve retrieval performance.

3. TAG FEATURES

The second modality we consider in this work are tags. While in the section above it is shown how to compute a topic model from visual features, we will now describe how to learn a high level representation for tag features by again using a pLSA model.

We assume in our work that all of the images in our database have been tagged by their authors, i.e. users have labeled their own images by specifying tags. Besides a single word, a tag can also be a phrase or a sentence. However in this work we treat each word of the images' annotations separately. Thus, in the following the term tag denotes a single word and is used interchangeably with "word".

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 times and by at least N_2 different users. This way all rarely used tags are neglected. We further filter the list by discarding all tag that contain a number and split tags at underscores into separate words. In a last filtering step all words within the vocabulary are checked whether they are known by Wordnet [7]. Wordnet is a lexical database of English. Only words that exist according to Wordnet build the final vocabulary.

Once the vocabulary is defined, a co-occurrence table is built by counting the tag occurrences for each image. However we expand the list of tags associated with an image by using Wordnet before forming the document vector. This is done to emphasize semantic features rather than the just simple word count features. We enrich the annotation of the image by the semantic parents of the user given tags.

Semantic parents for a word can easily be extracted from Wordnet, as in Wordnet each word is associated with hypernyms[1] and a words hypernyms denote its parents that express the specific concept of the tag more generally. Table 1 shows

---

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

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

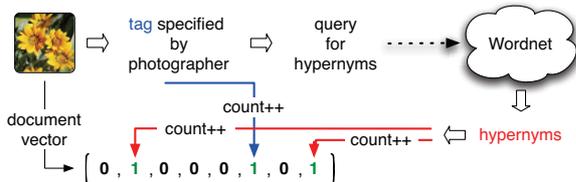**Table 1**. Examples for hypernyms (right) found in Wordnet for the words left.

Fig. 2. Building a document vector from tags and their hypernyms, assuming both tag and hypernyms are present in the vocabulary.

Fig. 3. Histogram of the total number of tags + hypernyms for each image within our database.

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 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 (Figure 2). In case the vocabulary does not contain a tag used for annotation, the word is simply ignored.

In our experiments we set the parameters for the tag vocabulary to $N_1 = 18$ and $N_2 = 10$ resulting in a vocabulary size of 2421. Most images in our database have between 5 and 15 tags associated (see Fig. 3). The number of tags for some images is however unreasonably large as users labeled images with whole sentences or phrases.

Once we have constructed co-occurrence vectors for each image we can use it to train a pLSA model and thus to build a high level image description based on tags. Again the topic distribution $P(z|d)$ is then used to obtain a representation for each image in some semantic tag space.

4. MULTIMODAL PLSA

Our proposed multimodal system fuses both modalities, visual features and tags, by learning a joint model for image representation. Therefore we compute a further pLSA model which is trained on top of the two base topic distributions, the one from the visual pLSA model and the one for the tag model. This higher-level topic model learns as well a topic distribution for each image, i.e. a "distribution over topic distributions" and thereby fuses the visual and tag representation. Moreover the topic model merges synonymous topics from different modalities and overcomes the problem of weighting the two modalities.
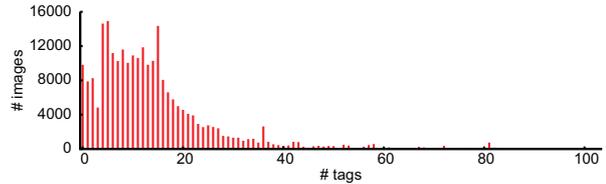
An overview of our complete multimodal pLSA system is

shown in Figure 4.

To train the third high-level pLSA model, the visual and tag representations given by the outputs of the two basic pLSA models are combined, i.e. we use as input the concatenated topic distributions of the training images learned by the two separate pLSA models on visual features and tags. Thus we use the high-level pLSA-based visual and tag features described in the previous section as input to our new pLSA model.

The topic mixture computed for each images with our multimodal pLSA is then used to represent each image in the database and we measure image similarity based on this representation.

5. EXPERIMENTAL EVALUATION

Our proposed retrieval system is experimentally evaluated on a dataset consisting of 246,347 Flickr images associated with at least a single tag [3]. We have downloaded geotagged images for 12 different categories and the database has not been cleaned or post-processed.

A visual vocabulary of size 10,000 is computed and we learn two 50 topic pLSA models, one for tag features and one for visual features. The model based on visual features has been trained with 50,000 images and the tag model has been learned from 10,000 images.

The multimodal pLSA model then maps the resulting 100-dimensional merged image representation (i.e. the two concatenated 50-dimensional topic vectors) to a multimodal topic distribution over 50 topics.

In our experiment, we compare our proposed approach to a system using topic mixtures based solely on visual features and a system which uses topic mixtures based on tag features for image representation. We use here the intermediate image descriptions we derive after applying our first two pLSA models. Note that the dimensionality of those vectors and the one of the multimodal representation are the same.

We evaluate the systems in a query-by-example task and judge the results in a user study. 60 query images are randomly selected and the L_1 distance is used to find their most similar images. The users are asked to rate the 19 closest results to each of our query images. **Note that we always show the images without their associated tags as we evaluate**
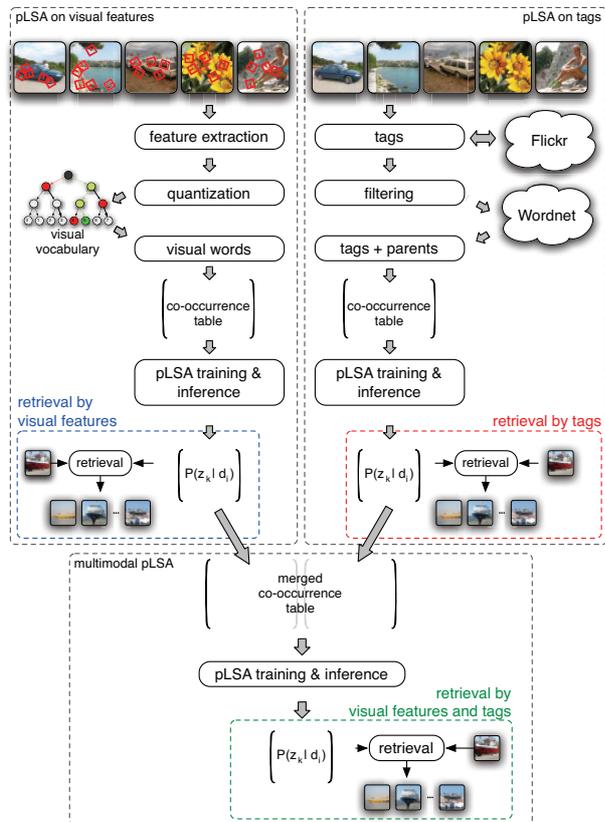
Fig. 4. Overview of our multimodal retrieval system.

Fig. 5. Scores for our different retrieval systems. Vertical bars mark the standard deviation between the users' means.

6. CONCLUSION

This work proposed a multimodal system exploiting both visual features and tags to derive a topic model that describes images. It is shown that the multimodal approach improves performance by approximately 36% compared to systems relying on visual or tag features alone. Extending or adding more modalities might further improve performance and is subject of future work.

7. REFERENCES

[1] Kobus Barnard, Pinar Duygulu, David Forsyth, David M Blei, Thomas Hofmann, Tomaso Poggio, and John Shawe-taylor, "Matching words and pictures," *Journal of Machine Learning Research 3 (2003) 1107–1135*, 2003.

[2] Thomas Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, Numbers 1-2, pp. 177–196, 2001.

[3] Rainer Lienhart and Malcolm Slaney, "pLSA on large scale image databases," *IEEE International Conference on Acoustics, Speech and Signal Processing 2007 (ICASSP 2007)*, vol. Vol. IV, pp. 1217–1220, 2007.

[4] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60 (2), pp. 91–110, 2004.

[5] David Nister and Henrik Stewenius, "Scalable recognition with a vocabulary tree," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2161–2168, 2006.

[6] Arthur Dempster, Nan Laird, and Donald Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[7] Christiane Fellbaum, *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

an query-by-image-example system. We use the following scoring to get a quantitative performance measure: An image considered being similar gets 1 point, an image considered as somewhat similar gets 0.5 points. All other images get 0 points. A mean score is calculated for each user and the mean over all users' means yields the final score of the system being evaluated.

The results of our experiment are shown in Figure 5. It can be seen that our proposed multimodal system outperforms the other two approaches. Furthermore the system using only tags shows improved performance over the system using visual features only. However, using both visual features and tag features and fusing those according to our proposed method increases the mean retrieval performance by about 36% over the tag-based system.

It should be noted that the variance of the means between different participants is quite large. This high variance might be an indicator for a very heterogeneous group of users. As the number of participants in our user-study is limited, outliers may have a deep impact on the overall score and were discarded. Finally, for each approach the ratings of 11 to 13 participants were u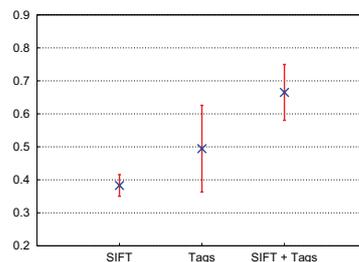sed to compute the final scores.