Modeling Semantic Aspects for Cross-Media Image Indexing

Florent Monay and Daniel Gatica-Perez, *Member*, *IEEE*

Abstract—To go beyond the query-by-example paradigm in image retrieval, there is a need for semantic indexing of large image collections for intuitive text-based image search. Different models have been proposed to learn the dependencies between the visual content of an image set and the associated text captions, then allowing for the automatic creation of semantic indexes for unannotated images. The task, however, remains unsolved. In this paper, we present three alternatives to learn a Probabilistic Latent Semantic Analysis (PLSA) model for annotated images and evaluate their respective performance for automatic image indexing. Under the PLSA assumptions, an image is modeled as a mixture of latent aspects that generates both image features and text captions, and we investigate three ways to learn the mixture of aspects. We also propose a more discriminative image representation than the traditional Blob histogram, concatenating quantized local color information and quantized local texture descriptors. The first learning procedure of a PLSA model for annotated images is a standard Expectation-Maximization (EM) algorithm, which implicitly assumes that the visual and the textual modalities can be treated equivalently. The other two models are based on an asymmetric PLSA learning, allowing to constrain the definition of the latent space on the visual or on the textual modality. We demonstrate that the textual modality is more appropriate to learn a semantically meaningful latent space, which translates into improved annotation performance. A comparison of our learning algorithms with respect to recent methods on a standard data set is presented, and a detailed evaluation of the performance shows the validity of our framework.

Index Terms—Image annotation, textual indexing, image retrieval, quantized local descriptors, latent aspect modeling.

1 Introduction

With the production of large digital image collections favored by cheap digital recording and storage devices, there is a clear need for efficient indexing and retrieval systems. The ideal system should allow for intuitive search for the user and require a minimal amount of human interaction to be applicable to large collections. Two distinct approaches to search large image collections coexist in the literature.

One is based on the *query-by-example* (QBE) paradigm [24], [32], [33], [7], [30]. In QBE systems, various low-level visual features are preliminarily extracted from the data set and stored as image indexes. The query is an image example that is indexed by its features, and retrieved images are ranked with respect to their similarity to this query index. Given that indexes are directly derived from the image content, this process requires no semantic labeling. The QBE paradigm is therefore an interesting solution for particular domains such as medical imaging [10], satellite images, and personal photo collections [13], where the query effectively exists as an image. These data sources tend to be specific, as the corresponding QBE solutions.

- *F. Monay is with the Signal Processing Institute, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland. E-mail: florent.monay@epfl.ch.*
- *D. Gatica-Perez is with the IDIAP Research Institute and École Polytechnique Fédérale de Lausanne, Centre du Parc, Av. des Pré-Beudin 20, Case Postale 592, CH-1920 Martigny, Switzerland. E-mail: gatica@idiap.ch.*

Manuscript received 1 Feb. 2006; revised 5 Nov. 2006; accepted 12 Dec. 2006; published online 19 Jan. 2007.
Recommended for acceptance by M. Hebert.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0091-0206.
Digital Object Identifier no. 10.1109/TPAMI.2007.1097.

QBE is not suitable for other types of image data sets. Commercial image collections such as Getty images or Corbis are searched with text-based queries because retrieval based on low-level visual similarity is, in general, not satisfactory for the user. The natural query process is in these cases textual and images in a collection are therefore indexed with words. Despite the development of systems and tools to assist it, this textual indexing process involves a substantial amount of work and usually results in heavy costs. Automatic image annotation has thus emerged as one of the key research areas in multimedia information retrieval, as an alternative to costly labor-intensive manual captioning [2], [3], [17], [12], [16], [18], [22], [9], [15], [26], [20].

Automatic image annotation systems take advantage of existing annotated image data sets to link the visual and textual modalities by using machine learning techniques. Although this framework seems very close to standard object detection [35], [1], key differences make automatic image annotation a distinct research problem. Although the vocabulary—the set of valid annotation words—might be constrained, captions from image collections can exhibit a large variability in general. Several words can describe one or more regions or even the whole image (see Fig. 1), which differs from the standard object detection scenario. Furthermore, the development of class-specific features and classifiers [34] is difficult, as the vocabulary size is usually much larger than the number of classes in standard object detection problems. Automatic image annotation systems therefore tend to rely on generic features and usually learn one model for the whole vocabulary [2], [3], [17], [12], [16], [22], [9], [15], [26], [20].

Independent of what features are chosen, the question is how to model the relation between captions and visual features to achieve the best textual indexing. A whole range of methods from a simple empirical distribution estimation to complex generative probabilistic models have been

flowers, garden, house, tower | buildings, people, sky, street | clouds, formation, sky, sunset

Fig. 1. Typical image captioning in the *Corel Stock Photo Library*.

proposed in the literature, offering a large variety of approaches. However, the difference in the nature of text captions and image features has not yet been fully investigated and exploited. In general, the textual and visual modalities are either considered as equivalent sources of data [22], [9], [20], [17], [26], or the caption words are simply considered as a class label [31], [8], [18] instead of a modality as such. The Correspondence Latent Dirichlet Allocation (CORR-LDA) [3] model is a notable exception, which builds a language-based correspondence between text and images. It first generates a set of hidden variables (latent aspects) that generate the regions of an image, decomposing an image into a mixture of latent aspects. A subset of these latent aspects is then selected to generate the text caption, which intuitively corresponds to the natural process of image annotation.

The CORR-LDA model acknowledges the complementarity of text and images as sources of information, as well as their difference in carrying semantic content, which needs to be taken into account to model the relation between modalities more accurately, with the goal of generating a better textual indexing. This paper investigates this concept, proposing a new dependence between words and image regions based on latent aspects. The contributions of our paper are the following. First, we present an alternative image representation to the standard Blob histogram, which combines quantized local color information and quantized local texture descriptors. Quantized versions of invariant local descriptors have been recently proposed as promising representations of objects and scenes [27], [11], [29] and applied to a small number of classes. However, to our knowledge, this representation has not been previously used in the context of image annotation, a more challenging problem from the number of concepts that is addressed. The effect of each type of visual features and their combination is analyzed in detail and we prove their complementarity by demonstrating the improvement of the retrieval performance for a majority of word queries for all the models that we consider. Second, we propose a probabilistic framework to analyze the contribution of the textual and the visual modalities separately. We assume that the two modalities share the same conditional probability distribution over a latent aspect variable that can be estimated from both or one of the two modalities for a given image. In this way, equal importance can be given to the visual and the textual features in defining the latent space or one of the two modalities can dominate. Based on extensive experiments, this framework allows us to show that the textual modality is more appropriate to learn a semantically meaningful latent space, which directly translates into an improved annotation performance. Finally, a comparison between different recently proposed methods is presented and a detailed evaluation of the performance shows the validity of our framework.

The paper is organized as follows: Section 2 presents an overview of the research in automatic image annotation and contrasts it with our work. Section 6.1 discusses the data and the visual representation considered in this work. Section 4 describes our probabilistic framework for image annotation. In Section 5, we discuss the state-of-the-art models that we implemented for comparison. Results and discussion are presented in Section 6.

2 Related Work

Existing works in automatic image annotation can be differentiated by the way in which they represent images and by the specific autoannotation model. These two aspects are used to guide the discussion in the following paragraphs.

A common first step to all automatic image annotation methods is the image segmentation into regions, using either a fixed-grid layout or an image segmentation algorithm. Regions have been described by a standard set of features including spatial frequencies, color, shape, and texture and handled as continuous vectors [2], [3], [17], [12], [16], [18] or in quantized form [22], [9], [15], [26], [20]. Different statistical assumptions about these quantized or continuous representations and image captions have led to different models. A representative selection of recent approaches is presented here.

The original approach described in [22] is based on a fixed-grid image segmentation and a vector quantization step. The color and texture representations of all training image blocks are quantized into a finite set of visual terms (*visterms*), which transforms an image into a set of visterms. All words attached to an image are attributed to its constituting visterms, and the empirical distribution of each word in the vocabulary given all visterms is computed from the set of training documents. A new image is indexed by first computing its visterms and then averaging the corresponding posterior distributions over words.

Contrary to [22], the work in [9] relies on the Normalized Cuts segmentation algorithm [28] to identify arbitrary image regions and build blobs. These blobs coarsely match objects or object parts, which naturally brings up a new assumption: The existence of an implicit one-to-one correspondence between blobs and words in the annotated image. The idea is borrowed from the machine translation literature and considers the word and blob modalities as two languages. An Expectation-Maximization (EM) procedure to estimate the probability distributions linking words and blobs is proposed. Once the model parameters are learned, words can be attached to a new image region. This *region naming* process is

comparable to object recognition, even if regions do not necessarily match objects, due to the obvious limitations of the segmentation algorithm. A new image is annotated by the most probable words given its constituting blobs.

The *cross-media relevance model* described in [15] also relies on a quantized blob image representation. However, unlike [9], it does not assume a one-to-one correspondence between blobs and words in images. Images are considered as sets of words and blobs, which are assumed independent given the image. The conditional probability of a word (respectively, blob) given a training image is estimated by the count of this word (respectively, blob) in this image smoothed by the average count of this word (respectively, blob) in the training set. These posterior distributions allow the estimation of the probability of a potential caption (set of words) and an unseen image (set of blobs) as an expectation over all training images. This annotation system improves the performance with respect to the machine translation method [9].

Linear-algebra-based methods applied on the word-by-document and Blob-by-document matrices are proposed in [26] to estimate the probability of a keyword given a blob. The correlation and the cosine measure between words and blobs are investigated to derive these conditional probabilities. The use of a Singular Value Decomposition (SVD) of the word-by-document and blob-by-document matrices, weighted with the term frequency–inverse document frequency (*tf-idf*) scheme, shows an improvement of the annotation performance over the original representation. A consistent improvement over the machine translation model [9] is shown.

In [17] and [12], Jeon et al. [15] abandon the quantization of image regions. With the same conditional independence assumptions as in their previous model [15], the continuous image region representation, modeled by a Gaussian Mixture Model (GMM), improves the image annotation performance. An additional modification is proposed in [12], where a multiple Bernoulli distribution for image captions replaces the multinomial distribution.

A statistical model of 600 image categories is proposed in [18]. Categories are labeled with multiple words, and images are manually classified in these categories. A 2D Multiresolution Hidden Markov Model (2D-MHMM) is learned on a fixed-grid segmentation of all category examples. The likelihood of a new image given each category's 2D-MHMM allows to select caption words for this image. This work is related to the *model vector indexing* approach [31], where one classifier (Support Vector Machine (SVM)) is trained for each semantic concept (seven concepts) and used for the indexing of a new image. The *Content-Based Soft Annotation* (CBSA) system [8] is also based on binary classifiers (Bayes Point Machines and SVMs) trained for each word (116 words are considered) and index a new image with the output of each classifier. The drawbacks are the learning of one classifier per word [31], [8] and of one model per set of words [18].

Different models to represent the joint distribution of words and image regions are discussed in [2], [3]. A hidden *aspect* variable is assumed in the data generative process, which links the textual and visual modalities through conditional relationships. This assumption translates into several variations of Latent Dirichlet Allocation (LDA)-based mixture models. Images are represented as a set of continuous region-based image features and modeled by Gaussian distributions conditioned on the aspects, whereas caption words are modeled with multinomial distributions. For instance, in the CORR-LDA model [3], words are conditioned on aspects that generated image regions. This additional constraint on word generation improves the overall annotation performance over less constrained LDA-based models.

A whole range of performance measures for automatic image annotation systems has been discussed in the literature. The quality of short image captions (≤ 5 words), intended to be similar to human annotations, has been evaluated with different measures [8], [9], [26], [15], [2], [20]. Specifically, the retrieval of images based on these short captions is evaluated with the precision and recall values of the retrieved image sets for a number of given queries in [31], [26], [9], [15]. Alternatively, the ratio of the correctly predicted words per image divided by the number of words in the ground-truth annotation has also been used for the evaluation of short captions [26], [2]. Proposed in [2], another measure for caption evaluation is the *Normalized Score*, which depends on the number of predicted words and allows to estimate the optimal number of words to predict [20]. However, the main goal of image annotation is to allow text-based queries for image retrieval, and this does not require the creation of short, fixed length text captions. All approaches (binary classification, probabilistic model, and linear algebra based) actually provide a confidence value for each word, which can be used for ranking all images in a collection. The confidence values for each word enable the construction of an image index that can be used for text-based image retrieval [31], [8], [15]. The average precision (AP) of a query (see Section 6.2), summarized by the mean average precision (mAP) for a set of queries, is then the natural metric for the retrieval performance. This way of annotating/indexing images and evaluating retrieval performance has started to become consensual [31], [15], and we therefore use it in this paper.

As it should be clear from this overview, the existing approaches proposed to learn relationships between visual and textual modalities in annotated images differ in the way images are represented, in the dependence assumptions that are made between image regions and words, and in the way model learning is performed. In this paper, we propose a probabilistic framework related to [2] and [3] that includes a hidden aspect variable to link the visual and textual modalities. This approach allows to consider regions and words from an image jointly, contrary to [22], where image regions are considered independently, and to [18] and [31], where categories (or words) are treated independently. Moreover, given that only one model is learned for all the words in the vocabulary, this type of approach might be better suited for large vocabularies than the supervised learning procedures proposed in [18], [31], which need to learn one model for each word. Finally, words and image features are of different nature and carry quite distinct levels of semantics and, so, we believe that these differences should influence how these two modalities are learned. Words and blobs are assumed equivalent in [9] (translation between two languages) and are treated equivalently in some of the models described in [2] and [3]. Unlike these works, we investigate different possibilities of learning the two modalities jointly while changing their respective influence.

In this sense, the closest work to ours is CORR-LDA, which first samples a latent aspect variable to generate an image region from a conditional Gaussian distribution and then samples an aspect from the same set of aspects to select a word from a conditional multinomial distribution. In contrast, in our work, we use multinomial distributions conditioned on aspects to model the discrete visual features, with the possibility to model a similar generative process as CORR-LDA, or to first generate the words and learn the

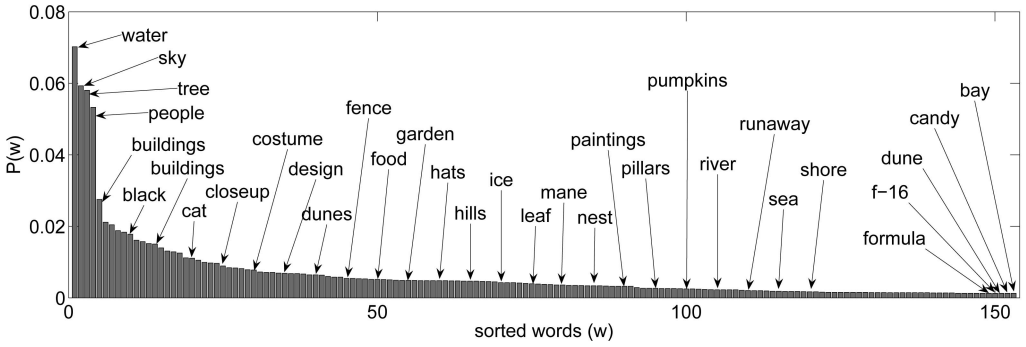

Fig. 2. Empirical distribution of words in the training images (training set 1, see Section 6.1). The most common words are water (1,124), sky (949), tree (929), people (853), and buildings (441). The least common words are formula (21), f-16 (21), dunes (21), candy (21), and bay (21). The numbers in brackets indicate the number of images in which each word occurs. Other words are shown to illustrate the nature of the vocabulary.

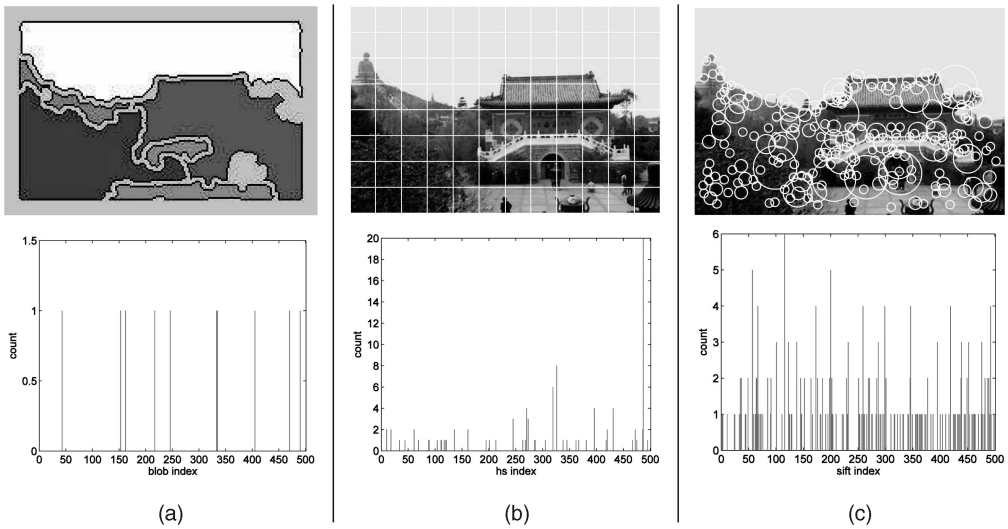

Fig. 3. Blob, HS, and SIFT image representations of the same image: (a) Normalized cut image segmentation from which texture, color, and shape features are extracted and the resulting histogram of the quantized image region features (Blobs), (b) uniform grid segmentation, from which color features are extracted, and the resulting histogram of the quantized image region features (HS), and (c) regions detected by the DoG point detector and the resulting histogram of the quantized SIFT descriptors.

related aspect distributions that we later link to the visual features. As stated in the introduction, we also propose an enriched image representation that includes quantized local image descriptors that has not been investigated in autoannotation but used in very recent works for scene and object classification [11], [29], [27]. We conduct a thorough study comparing various competitive methods with a consistent evaluation procedure and show an improvement of performance.

3 Annotated Image Representation

3.1 Text Caption Representation

Images in our data set are annotated with a few unordered words selected from a vocabulary of size N_w. The representation of the caption of an image d_i is a histogram $w(d_i)$ of size N_w

$$w(d_i) = \{n(d_i, w_1), \ldots, n(d_i, w_j), \ldots, n(d_i, w_{N_w})\}, \quad (1)$$

where $n(d_i, w_j)$ denotes the count of the word w_j in the caption of the image d_i. This is a standard representation for text documents, which could also be used in the case of free-text captions after the word stopping and word stemming preprocessing steps. As shown in Fig. 2, the distribution of words is highly skewed. As the data set mainly consists of outdoor images, the words *water*, *sky*, *tree*, *people*, and *buildings* account for a big proportion of the probability mass. The empirical distribution also shows that there are many words represented by only a few training examples that nevertheless will have to be predicted, which advocates for a model that learns the co-occurrence of these infrequent words with more frequent words in order to predict them with higher accuracy. Training a separate model for a specific infrequent word seems difficult, whereas identifying the words with which this word co-occurs could be, instead, a good strategy.

3.2 Image Representation

We investigate three types of image representations based on quantized image regions that are illustrated on Fig. 3. The first relies on large-scale image regions, combining both texture and color information (see Fig. 3a). The two other image representations are based on a larger number of smaller scale image regions, uniformly extracted from a fixed grid (see Fig. 3b) or identified by a point detector (see Fig. 3c). They capture color or texture information, respectively. The three discrete feature types are described in the following.

3.2.1 Blobs

We consider an image representation originally proposed for region-based QBE [7] and later used for image annotation [2], [9], [3], [15]. A maximum of 10 regions per image, identified by the normalized cut segmentation algorithm [28], are represented by 36 features including color (18), texture (12), and shape/location (6). The K-means clustering algorithm is then applied to the region descriptors, quantizing them into a N_b-dimensional *Blob* representation. Note that the difference in the number of feature components makes the resulting Blob representation intrinsically biased toward color. An image d_i is segmented into a set of large image regions that are quantized and represented by the corresponding histogram $b(d_i)$ of size N_b (see Fig. 3a)

$$b(d_i) = \{n(d_i, b_1), \ldots, n(d_i, b_j), \ldots, n(d_i, b_{N_b})\}, \quad (2)$$

where $n(d_i, b_j)$ denotes the number of regions in image d_i that are quantized into the Blob b_j. The motivation behind this representation is a possible match between the automatically segmented image regions and objects in the images. We see, for instance, in Fig. 3a that the green region matches trees in the original image and that the sky is covered by exactly one blob. As mentioned in [12], the match between the segmented regions and objects in the image is, however, relatively poor in general.

3.2.2 Hue Saturation (HS)

In spite of progress, no automatic segmentation algorithm is currently capable of dividing an image into consistently meaningful parts. The use of a segmentation algorithm is therefore difficult to justify and we decided to extract image regions from a uniform grid, as illustrated in Fig. 3b. The pixel color distribution from the resulting regions is represented by a 2D HS histogram, where the color brightness value from the Hue-Saturation-Value (HSV) color space is discarded for illumination invariance [25]. These HS histograms are then quantized into N_h bins with the K-means clustering algorithm to obtain the corresponding histogram representation $h(d_i)$ of size N_h for the image d_i (see Fig. 3b)

$$h(d_i) = \{n(d_i, h_1), \ldots, n(d_i, h_j), \ldots, n(d_i, h_{N_h})\}, \quad (3)$$

where $n(d_i, h_j)$ denotes the number of regions in image d_i that are quantized into the HS bin h_j. Contrary to a global color histogram, $h(d_i)$ encodes the distribution of color information from local image regions. In the rest of the paper, we refer to this representation of an image as the HS representation.

3.2.3 Scale-Invariant Feature Transform (SIFT)

We also propose the use of a third image representation based on local descriptors computed over automatically sampled image regions (see Fig. 3c). Very recently, these features have been successfully combined with probabilistic latent space models [11], [29], [27] and have shown good performance in modeling different types of image content, including objects [29], [21] and scenes [11], [27], to, roughly speaking, a dozen concepts. However, its applicability to a much larger number of semantic concepts, to our knowledge, has not been investigated. In this representation, the image is first sampled with the difference-of-Gaussians (DoG) point detector [19] at different scales and locations (see Fig. 3c). This detector has been built to be invariant to translation, scale, rotation, and illumination changes and samples images at different locations and scales, depending on their content. We represent each detected region with the SIFT descriptor, which consists of a histogram of edge directions at different locations [19]. The SIFT descriptors are then quantized by the K-means clustering algorithm to obtain a discrete set of local N_s image-patch indexes. An image d_i is represented by the histogram $s(d_i)$ of its constituting local patches (see Fig. 3c)

$$s(d_i) = \{n(d_i, s_1), \ldots, n(d_i, s_j), \ldots, n(d_i, s_{N_s})\}, \quad (4)$$

where $n(d_i, s_j)$ is the number of local descriptors in the image d_i that have been quantized into the image patch s_j. In the rest of the paper, we refer to this representation as the SIFT representation.

The Blob, HS, and SIFT image representations encode different image properties and are therefore expected to produce different performances. The Blob representation is based on the joint quantization of shape, texture, and color features, extracted from large image regions. The HS and SIFT representations are respectively based on the quantization of color and texture information, extracted from small-scale image regions. As shown in Fig. 3, the number of regions that are considered in each case also varies: A maximum of 10 regions per image in the Blob case, 96 32-by-32-pixel square regions in the HS case and an average of 240 detected points (depending on the image content) in the SIFT case. This makes the Blob histogram sparser than the HS and SIFT histograms for an equivalent number of 500 K-means clusters, as shown in Figs. 3a, 3b, and 3c. In Section 6, we investigate the combination of these image representations. For instance, using a direct concatenation of them in a first evaluation, the concatenation of the HS and SIFT features forms the complementary $v(d_i) = \{h(d_i), s(d_i)\}$ histogram of size $N_v = N_h + N_s$. To take advantage of these complementary sources of visual information, the methods have to treat these unbalanced representations efficiently.

4 Linked Aspect Models for Annotated Images

4.1 Aspect Models for Text

An intuitive way to describe a text document is by considering the different topics it consists of. These topics are not explicit but can be derived from the documents themselves and represent an accurate and compact summary of the original content. People usually compare text documents based on their respective topic distribution and do not tend to evaluate the similarity directly at the word level (unless looking for duplicates). To achieve a semantically meaningful indexing of texts, an increasing body of research in information retrieval aims at discovering methods to automatically identify hidden topics in a set of text documents.

Different *latent aspect models* [14], [6], [4] have been proposed to achieve this task. Their common assumption is the sampling of a *hidden variable* (referred to as *latent aspect*) in the generative process of words in a document. Documents from the same text corpora share these latent aspects: a document is a mixture of latent aspects. These latent aspects are defined by multinomial distributions over words that are learned for each text corpus considered. These distributions characterize the aspects and show that a correspondence between topics identified by humans and latent aspects can exist [14], [4].

Model parameters are estimated from the co-occurrence of words across documents and learning is therefore fully

unsupervised. The distribution over aspects represents an alternative text document representation and can improve the classification of text documents into categories [4] or provide a less ambiguous representation for text retrieval applications [14].

4.2 Aspect Models for Annotated Images

The concept of latent aspects is not restricted to text documents. Images are intuitively seen as mixtures of several content types, which make them good candidates for a latent aspect approach. Different latent aspect models, adapted from the LDA [4] model for text, have been proposed to model annotated images [2], [3]. A Gaussian distribution on the continuous visual feature space models the visual modality for each aspect, and different ways to combine the textual and the visual modalities using latent aspects are discussed. In particular, as presented in Section 2, the CORR-LDA model [3] shares the latent aspects between the visual and textual modalities by first generating the image features and then generating the words from the subset of aspects that generated the image features. In the CORR-LDA model, the visual modality thus drives the definition of the latent space to which the textual modality is then linked.

The CORR-LDA model shows that driving the definition of the latent aspects by the visual modality and then sharing these aspects with the text modality is more appropriate than an unspecified loose dependence between the two modalities. In our work, we investigate this key concept in more detail and compare the effect of a latent space driven by the visual features with a latent space driven by the textual features. We also compare the performance of these two options with the performance of a latent space learned from a concatenation of the two modalities. To conduct this analysis, three alternative procedures to learn a Probabilistic Latent Semantic Analysis (PLSA) model [14] for annotated images are proposed in Section 4.4. Although the LDA model has been shown to improve over PLSA in terms of perplexity [3] in text collections, we have chosen to base our investigation on PLSA since it allows for an exact EM algorithm. This makes the intended modifications of the learning procedure easier, without harming the resulting analysis. Moreover, PLSA has been recently shown to perform well on image classification tasks [27], [29], using the aspect mixture proportions to learn the classifiers. In particular, Sivic et al. [29] recently compared the PLSA and the LDA models for image classification and they showed that a higher classification performance was obtained with PLSA in that case. We describe the PLSA model in Section 4.3 and our proposed three alternative methods to model annotated images with PLSA in Section 4.4.

4.3 The PLSA Model

PLSA [14] assumes the existence of a latent aspect z_k ($k \in 1, \ldots, N_z$) in the generative process of each element x_j ($j \in 1, \ldots, N_x$) in a document d_i ($i \in 1, \ldots, N_d$). Each occurrence x_j is independent from the document it belongs to, given the latent variable z_k, which corresponds to the joint probability expressed by

$$P(x_j, z_k, d_i) = P(d_i)P(z_k \mid d_i)P(x_j \mid z_k). \quad (5)$$

The joint probability of the observed variables is the marginalization over the N_z latent aspects z_k as expressed by

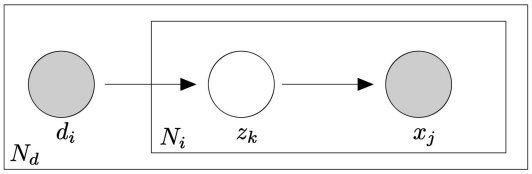

Fig. 4. The PLSA graphical model. Shaded nodes are observed. For each observation pair (d_i, x_j), x_j is independent of d_i, given the latent aspect z_k. N_d is the number of training documents d_i and N_i is the number of elements x_j in d_i.

$$P(x_j, d_i) = P(d_i) \sum_{k=1}^{N_z} P(z_k \mid d_i)P(x_j \mid z_k). \quad (6)$$

The graphical model shown in Fig. 4 illustrates the conditional independence assumptions of the PLSA model expressed in (5). A document d_i is first selected with the probability $P(d_i)$, which is proportional to the size of the document d_i, and an aspect z_k is selected from the conditional probability distribution $P(z|d_i)$. Given the aspect z_k, an element x_j is selected according to the conditional probability distribution $P(x|z_k)$. More details of the model are presented in the following paragraphs.

4.3.1 Model Parameters

The conditional probability distributions $P(x|z_k)$ and $P(z|d_i)$ are multinomial, given that both z and x are discrete random variables. The parameters of these distributions are estimated by the EM algorithm [14]. For a document collection containing N_x different elements, $P(x|z)$ is an N_x-by-N_z table that stores the parameters of the N_z multinomial distributions $P(x|z_k)$. $P(x|z_k)$ characterizes each aspect z_k and is valid for documents that are not part of the training set. On the contrary, the N_z-by-N_d table $P(z|d)$ is only relative to the N_d training documents, as it stores the parameters of the N_d multinomial distributions $P(z|d_i)$ that describe the training document d_i.

To illustrate these conditional probability distributions in the context of image captions, Fig. 5c shows the PLSA decomposition of an image caption in $N_z = 80$ aspects (N_z chosen arbitrarily), where the parameters are learned on the captions of 5,188 images. The PLSA model decomposes the caption into three main aspects, which are represented in Figs. 5d, 5e, and 5f by their multinomial distributions over words $P(w|z_k)$. As can be seen, aspect number 10 (Fig. 5d) is most likely to generate the word *mountain* (then, *valley*); aspect number 3 (Fig. 5e) generates the words *temple, statues, sculpture,* and *shrine*, with high probabilities; and aspect 47 (Fig. 5f) is related to the words *stone, ruins, sculpture, pillars*, and *pyramid*.

4.3.2 Learning

An EM algorithm can be derived from the likelihood of the observed data (7) to estimate the parameters of the distributions $P(x|z)$ and $P(z|d)$

$$\mathcal{L} = \prod_{i=1}^{N_d} \prod_{j=1}^{N_x} P(d_i) \sum_{k=1}^{N_z} P(z_k \mid d_i)P(x_j \mid z_k)^{n(d_i, x_j)}, \quad (7)$$

where $n(d_i, x_j)$ is the count of element x_j in document d_i. The two steps of the EM algorithm are described as follows:

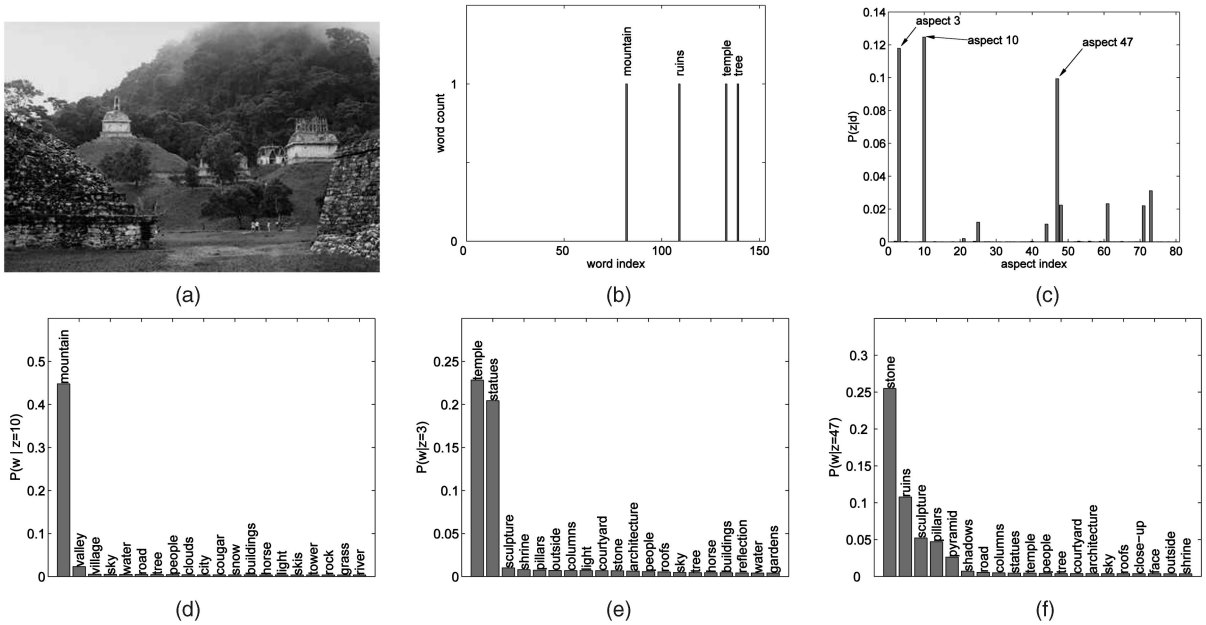

Fig. 5. Aspect decomposition of the image captions *mountain, ruins, temple*, and *tree* for a PLSA model trained with 80 aspects on 5,188 image captions. (a) is the considered image d_i, (b) is the word caption histogram $w(d_i)$, (c) is the aspect distribution $P(z|d_i)$, and (d) and (f) are the distributions of the 20 top-ranked words given the three most probable aspects (10, 3, and 47, respectively).

- **E-step**. The conditional probability distribution of the latent aspect z_k given the observation pair (d_i, x_j) is computed from the previous estimate of the model parameters:

$$P(z_k \mid d_i, x_j) = \frac{P(x_j \mid z_k)P(z_k \mid d_i)}{\sum_{k=1}^{N_z} P(x_j \mid z_k)P(z_k \mid d_i)}. \quad (8)$$

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

$$P(x_j \mid z_k) = \frac{\sum_{i=1}^{N_d} n(d_i, x_j)P(z_k \mid d_i, x_j)}{\sum_{m=1}^{N_x}\sum_{i=1}^{N_d} n(d_i, x_m)P(z_k \mid d_i, x_m)}, \quad (9)$$

$$P(z_k \mid d_i) = \frac{\sum_{j=1}^{N_x} n(d_i, x_j)P(z_k \mid d_i, x_j)}{n(d_i)}. \quad (10)$$

4.3.3 Inference: PLSA of a New Document

The conditional probability distribution over aspects $P(z|d_{new})$ can be inferred for an unseen document d_{new}. The *folding-in* method proposed in [14] maximizes the likelihood of the document d_{new} with a partial version of the EM algorithm described above, where $P(x|z)$ is obtained from training and *kept fixed* (that is, not updated at each M-step). In doing so, $P(z|d_{new})$ maximizes the likelihood of the document d_{new} with respect to the previously learned $P(x|z)$ parameters.

4.3.4 Overfitting Control

We control the overfitting of the model by early stopping, based on the likelihood of a validation set. We consider the *folding-in likelihood*, which allows good performance prediction and overfitting control without the need for a tempered version of the EM algorithm [5]. The probability of aspects given each validation document $P(z|d_{valid})$ is first estimated using the folding-in method, as described in Section 4.3.3. The folding-in likelihood of the validation set given the current parameters is then defined as

$$\mathcal{L}_{valid} = \prod_{i=1}^{N_{d_{valid}}}\prod_{j=1}^{N_x} P(x_j \mid d_i) = \prod_{i=1}^{N_{d_{valid}}}\prod_{j=1}^{N_x}\sum_{k=1}^{N_z} P(x_j \mid z_k)P(z_k \mid d_i). \quad (11)$$

4.4 Modeling Annotated Images with PLSA

We discuss here three alternatives to learn a PLSA model for the co-occurrence of visual and textual features in annotated images. The first is a direct application of PLSA to the early integration of visual and textual modalities [20]. The two others are based on a variation of the PLSA EM algorithm that constrains the estimation of the conditional distributions of latent aspects given the training documents from one of the two modalities only. This allows choosing between the textual and the visual modality to estimate the mixture of aspects in a given document, which constrains the definition of the latent aspects on one or the other modality.

4.4.1 PLSA-MIXED

The PLSA-MIXED [20] model learns a standard PLSA model on a concatenated representation of the textual and the visual features $x = (w, v)$. Using a training set of captioned images, $P(x|z)$ is learned for both textual and visual co-occurrences to capture simultaneous occurrence of visual features and words. Once $P(x|z)$ has been learned, it can be used to infer a distribution over words for a new image as follows: The new image d_{new} is represented in the concatenated vector-space, where all word elements are zero (no annotation): $x_{new} = (0, v_{new})$. The multinomial distribution over aspects given the new image $P(z|d_{new})$ is then computed with the partial PLSA steps described in

Section 4.3 and allows the computation of $P(x|d_{new})$. The conditional probability distribution over words $P(w|d_{new})$ is extracted from $P(x|d_{new})$ and allows the annotation of the new image d_{new}.

4.4.2 Asymmetric PLSA Learning

We propose to model a set of annotated images with a PLSA model for which the conditional distributions over aspects $P(z|d_i)$ is estimated from one of the two modality only. Contrary to PLSA-MIXED, which learns $P(z|d_i)$ from both the visual and the textual modalities, this formulation allows to treat each modality differently, giving more importance to the text captions or the image features in the latent space definition. We refer to this alternative learning algorithm as an *asymmetric PLSA learning*. Intrinsically, PLSA-MIXED assumes that the two modalities have an equivalent importance in defining the latent space, given that the latent space is learned from their concatenated representation. The only potential imbalance could result from a marked difference between the number of words and the number of visual features in the images, and these values are not freely controlled in practice.

An asymmetric PLSA learning gives a better control of the respective influence of each modality in the latent space definition. This concretely allows modeling of an image as a mixture of latent aspects that is defined either by its text captions or by its visual features, resulting in different mixtures. The aspect distributions $P(z|d_i)$ are learned for all training documents from one modality only (visual or textual modality) and are kept fixed for the other modality (textual or visual modality, respectively). We refer to PLSA-FEATURES when the aspect distributions are learned on the visual features and to PLSA-WORDS when the aspect distributions are learned on the image captions. In the following, we describe how the parameters are learned in the asymmetric learning case and how the distribution over words is estimated for a new document.

Learning parameters. The description of the learning process is valid for the PLSA-FEATURES and the PLSA-WORDS approaches but differs on which modality the multinomial distribution over aspects are learned for the training documents. The first and second modalities are therefore referred to as x^1 and x^2, respectively, and correspond either to the visual or to the textual features in the following:

1. The first modality is used to estimate the N_z conditional distributions $P(x^1|z_k)$ and the N_d conditional distributions $P(z|d_i)$ on the training set.
2. We consider that the aspects have been observed for the set of training documents and estimate the N_z conditional probability distributions $P(x^2|z_k)$ for the second modality, keeping $P(z|d)$ from above fixed. Note that this technique is computationally similar to the standard PLSA procedure for inference in unseen documents described in Section 4.3, where $P(x|z)$ is kept fixed and $P(z|d)$ is computed by likelihood maximization. However, what we are trying to do is, conceptually speaking, very different.

The parameters of the $P(x^1|z_k)$ and $P(x^2|z_k)$ distributions are defining the latent aspects z_k based on the visual and textual modalities, respectively, for PLSA-FEATURES, and conversely, for PLSA-WORDS. Early stopping is performed for each of the two learning steps described above. In the second step, the probability of aspects given the validation documents $P(z|d_{valid})$ estimated from the first step are not reestimated by folding-in.

Annotation by inference. Given new visual features $v(d_{new})$ and the previously estimated $P(v|z)$ parameters, the conditional probability distribution $P(z|d_{new})$ is inferred for a new image d_{new} using the standard PLSA procedure for a new document (Section 4.3). Furthermore, the conditional distribution of the words given this new image is given by

$$P(w|d_{new}) = \sum_{k=1}^{N_z} P(w \mid z_k) * P(z_k \mid d_{new}). \tag{12}$$

5 Baseline Methods

Three baseline models for image annotation are considered for comparison with our models. The first baseline consists in a visual comparison between the image to annotate and the training images, propagating their annotations based on this similarity. The two other methods correspond to the state-of-the-art performance in image annotation when the discrete quantized Blob representation $b(d)$ is used [9], [15], [26], [20].

5.1 Annotation Propagation

Intuitively, training images that are similar to a new image d_{new} should be taken into account to generate its annotation. Our simplest baseline therefore consists in computing the similarity between the image d_{new} and the training images, sequentially attaching their respective annotation to d_{new} based on these similarities. Concretely, we compute the cosine similarity between the image d_{new} and the N_d training images d_i based on their respective visual representations $v(d_{new})$ and $v(d_i)$,

$$sim_{cos}(v(d_{new}), v(d_i)) = \frac{\sum_{j=1}^{N_v} n(d_{new}, v_j) n(d_i, v_j)}{\sqrt{\sum_{j=1}^{N_v} n(d_{new}, v_j)^2} \sqrt{\sum_{j=1}^{N_v} n(d_i, v_j)^2}}. \tag{13}$$

The training images are ranked with respect to this similarity measure and the probability of a word w_i given d_{new} is estimated by the inverse of the best ranked image according to (13) that contains the word w_i

$$P(w_i \mid d_{new}) \propto (rank(d_{best}))^{-1}, \tag{14}$$

where d_{best} is the most similar image to d_{new} in the training set that contains the word w_i and $rank(d_{best})$ is the rank order of this image given d_{new}. The word probabilities are then normalized so that $\sum_{i=1}^{N_w} P(w_i|d_{new}) = 1$.

5.2 Cross-Media Relevance Model

In [15], the annotation of an unseen image d_{new} is based on the joint probability of all its m constituting visual elements v_l and the word w_j. This joint probability is estimated by its expectation over the N_d training images

$$P(w_j, v_1, \ldots, v_m) = \sum_{i=1}^{N_d} P(d_i) P(w_j, v_1, \ldots, v_m \mid d_i). \tag{15}$$

The visual elements are considered independent given an image d_i, which gives

$$P(w_j, v_1, \ldots, v_m) = \sum_{i=1}^{N_d} P(d_i) P(w_j \mid d_i) \prod_{l=1}^{N_v} P(v_l \mid d_i)^{n(d_i, v_l)}, \quad (16)$$

where $n(d_i, v_l)$ is the count of the visual element v_l in the image d_i. The probability of a word w in a training image d_i is the likelihood of this word in this image combined with the likelihood of this word in all the training images. A fusion parameter α controls the importance of the image and the training set likelihoods

$$P(w_j \mid d_i) = (1-\alpha) \frac{n(d_i, w_j)}{\sum_{l=1}^{N_v} n(d_i, v_l) + \sum_{j=1}^{N_w} n(d_i, w_j)} + \alpha \frac{n(w_j, d)}{N_d}, \quad (17)$$

where $n(d_i, w_j)$ denotes the count of the word w_j in the image d_i, $n(d_i, v_l)$ is the count of the visual element v_l in the image d_i, $n(w_j, d)$ is the number of images in which the word w_j appears, and N_d is the number of training images. Similarly, the probability of a visual element given an image d_i is estimated by its likelihood in this image smoothed by its likelihood in the training set, controlled by a parameter β

$$P(v_l \mid d_i) = (1-\beta) \frac{n(d_i, v_l)}{\sum_{l=1}^{N_v} n(d_i, v_l) + \sum_{j=1}^{N_w} n(d_i, w_j)} + \beta \frac{n(v_l, d)}{N_d}, \quad (18)$$

where $n(v_l, d)$ is the number of images in which the visual element v_l appears. The parameters α and β are estimated on a validation set to optimize the model performance.

5.3 Cross-Media Translation Table

In [26], a translation table T_{cos} between words and quantized visual features is proposed. The word-by-image matrix is weighted with the *tf-idf* scheme to obtain the weighted matrix D_w

$$D_w = \left(n(d_i, w_j) * log\left(\frac{N_d}{n(w_j, d)} \right) \right)_{N_d \times N_w}, \quad (19)$$

where $n(d_i, w_j)$ is the count of the word w_j in the image d_i, $n(w_j, d)$ is the number of documents the word w_j appears in, N_d is the number of training images, and N_w is the size of the vocabulary. Similarly, the feature-by-image matrix is weighted with the *tf-idf* scheme to obtain the weighted matrix D_v

$$D_v = \left(n(d_i, v_l) * log\left(\frac{N_d}{n(v_l, d)} \right) \right)_{N_d \times N_v}, \quad (20)$$

where again $n(d_i, v_l)$ is the count of the visual element v_l in the image d_i, $n(v_l, d)$ is the number of documents the visual element v_j appears in, and N_v is the size of the visual feature space. An SVD is applied on the D_w and D_v matrices, keeping the first r eigenvalues that preserve 90 percent of the variance to suppress the noise in the data. Let the jth column of the matrix D_w be d_{wj} and the lth column of the matrix D_v be d_{vl}. The cross-media translation table T is defined by

$$T_{cos} = (sim_{cos}(d_{wj}, d_{vl}))_{N_w \times N_v}, \quad (21)$$

where the cosine similarity function $sim_{cos}()$ is defined in (13). Normalizing T_{cos} by column, the annotation of a new image d_{new} represented by its histogram $v(d_{new})$ is given by

$$P(w \mid d_{new}) = T_{cos} \times v(d_{new}). \quad (22)$$

6 Results

6.1 Data

As shown in [23], for the case of QBE, contradictory rankings can be obtained if the performance evaluation is conducted on different data subsets, even if these subsets are created from the same original image collection. To prevent this possible inaccuracy, it is crucial to compare different systems on identical data, with clearly defined training and testing sets. We conduct our experiments on an annotated image data set that was originally used in [2] and consists of 10 samples of roughly 16,000 annotated images. Each sample is split into a training set and a testing set, with an average number of 5,240 training and 1,750 testing images. The average vocabulary size is 161. The Blob representation, as well as the description of the different samples, was downloaded from http://kobus.ca/research/data/jmlr_2003/.

6.2 MAP Measure

A number of works [9], [26], [12], [16], [15] measure the ability of the system to produce a human-like annotation, selecting a small number of words from the vocabulary. A fixed threshold or a fixed number of words has to be decided to extract short captions that can be used for image retrieval. With this, for a given query word, the number of correctly retrieved images divided by the number of retrieved images is the *word precision*, and the number of correctly retrieved images divided by the total number of correct images is the *word recall*. The average word precision and word recall values summarize the system performance.

One drawback of creating a human-like annotation is that only a fraction of words from the vocabulary are eventually predicted for the test images, because uncommon words tend not to be predicted due to a very low conditional probability. The word precision and recall values have thus to be presented together with the number of predictable words, as done in [9], [12], [16], [15], which makes the comparison between models unclear. Is it better for a system to predict only a few words with a high accuracy or is a system more efficient if it can predict more words?

However, given that the goal is to index images for image retrieval, there is no need to produce such short human-like annotation. The conditional probability distribution $P(w|d_{new})$ can be used to rank the images for all possible queries. Even if the conditional probabilities of a specific word might be low for some of the images, the comparison of the relative probability values allows ranking of the image collection for each word query. To illustrate this, the truncated word distribution inferred from two images using the PLSA-WORDS model is shown in Fig. 6. The word *flowers* is in the top 20 words for both images, and the probability of the word *flowers* given the top image is higher than given the bottom image. This information would be discarded if the model is used to predict a fixed length annotation, although it can obviously be exploited for image ranking. The distribution over words in Fig. 6 also shows how much more probable the

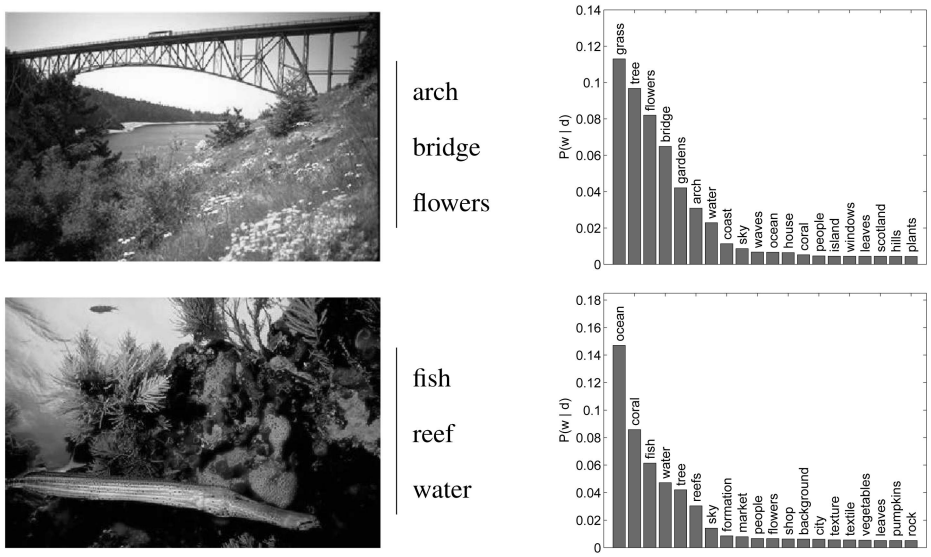

Fig. 6. The conditional probability distribution $P(w|d)$ inferred from two test images from their HS+SIFT representation with the PLSA-WORDS approach. The image and the ground-truth annotation are shown in (a) and the top 20 words and their conditional probability are shown in (b).

TABLE 1
Average mAP Values (in Percent) over 10 Cross-Validation Runs for Different Quantization of the HS and the SIFT Image Representations for the Three Baseline Methods

	Blobs	HS				SIFT			
	500	100	200	500	1000	100	200	500	1000
propagation	10.2 (0.6)	10.7 (0.7)	10.8 (0.6)	11.7 (0.7)	12.4 (1.0)	10.7 (0.9)	11.4 (0.6)	12.2 (0.8)	13.0 (0.8)
CMRM [15]	12.1 (0.8)	13.4 (1.0)	14.2 (0.9)	14.4 (1.1)	14.5 (1.2)	11.6 (0.9)	12.7 (1.0)	12.3 (1.6)	10.0 (2.0)
SVD-cos [26]	15.6 (0.7)	14.1 (1.0)	15.4 (1.0)	16.4 (1.1)	17.1 (1.1)	10.0 (0.8)	11.6 (0.8)	12.8 (0.9)	14.3 (0.9)

The standard deviation is given in parentheses.

word *ocean* is given the bottom image than given the top-image. This, however, would not be possible if we were only relying on a five-word annotation.

The performance measure used in this work is mean average precision. This has been a standard measure for the retrieval of text documents for years, which has also been used by Text Retrieval Conference Video Retrieval Evaluation (TRECVID) to evaluate the semantic concept video retrieval task for several years (details can be found at http://www-nlpir.nist.gov/projects/trecvid/). mAP has the ability to summarize the performance in a meaningful way. To compute it, the AP of a query q is first defined as the sum of the precisions of the correctly retrieved images at rank i divided by the total number of relevant images $rel(q)$ for this query

$$mAP = \frac{\sum_{q=1}^{N_q} AP(q)}{N_q}, \quad \text{where}$$
$$AP(q) = \frac{\sum_{i \in relevant} precision(i)}{rel(q)}. \tag{23}$$

The AP measure of a query is thus sensitive to the entire ranking of documents. The mean of the AP of N_q queries summarizes the performance of a retrieval system in one mAP value.

6.3 Hyperparameters and Cross Validation

We need to estimate two types of hyperparameters by cross validation. The first is the number of K-means clusters that defines the quantization of the visual features, and the second is the number of latent aspects for the approaches based on a PLSA model. The number of K-means clusters is cross-validated for the HS, SIFT, and HS+SIFT representations for 100, 200, 500, and 1,000 clusters. The value of $N_b = 500$ clusters for the Blob representation is kept fixed, as this representation is provided as is by Barnard et al. [2]. The mAP performance of the Blob representation is given for comparison. The K-means models are learned on the training images of each sample set. In Table 1, we show the mAP values obtained with the three baseline methods (propagation, Cross-Media Relevance Model (CMRM), and SVD-COS), averaged over 10 cross-validation runs for one of the 10 sample sets. The hyperparameter values estimated by cross validation from this sample set will be used for the remaining nine, as one set is assumed to be representative of the entire set. For the three baseline methods, the best number of K-means clusters for both HS and SIFT representations is 1,000, except for the SIFT representation in the CMRM case for which 200 clusters corresponds to the best retrieval performance. We also observe that the HS representation consistently achieves a higher performance than the Blob representation for the same number of

TABLE 2
Average mAP Values (in Percent) over 10 Cross-Validation Runs for Representations Based on the Concatenation of Different Quantization of the HS and SIFT Features for the Three Baseline Methods

	HS+SIFT			
	500-500	500-1000	1000-500	1000-1000
propagation	16.0 (1.3)	15.5 (1.4)	16.8 (1.2)	16.4 (1.3)
CMRM [15]	17.6 (0.8)	17.4 (0.8)	6.2 (0.8)	4.8 (0.8)
SVD-cos [26]	19.9 (1.5)	20.9 (1.7)	20.2 (1.7)	21.2 (1.7)

The standard deviation is given in parentheses.

clusters. We use these estimated number of clusters for the final performance evaluation in Section 6.4.

We also estimated the number of clusters by cross validation for the HS+SIFT concatenation, as reported in Table 2. We restricted our analysis to the combination of HS and SIFT features for two reasons. First, as Table 1 suggests, the HS representation outperforms the Blob representation. Second, HS and SIFT features result from the quantization of local-color-only and local-texture-only information, respectively, whereas the Blob representation corresponds to the joint quantization of color, texture, and shape. Analyzing the effect of the combination of separately extracted color-only and texture-only information seems more intuitive than analyzing the combination of a texture-based representation with a joint color-texture-shape representation. The values in Table 2 show that the optimal (N_h, N_s) HS-SIFT combination for the propagation method is (1,000, 500), (500, 500) for the CMRM case and (1,000, 1,000) for the SVD-COS case. The results from the CMRM method drop significantly for the (1,000, 500) and (1,000, 1,000) combination, although we carefully selected the α and β parameters.

Regarding the PLSA-based approaches, as we have mentioned, they require the number of latent aspects N_z to be estimated, as this hyperparmeter defines the capacity of the model: the number of parameters $P = (N_d(N_z - 1)) + (N_z(N_x - 1))) \sim N_z(N_d + N_x)$ linearly depends on N_z. The best value for the number of clusters therefore needs to be jointly estimated with the number of latent aspects for the three PLSA-based approaches, which is presented in Fig. 7. The average of the mAP values computed for 10 cross-validation runs are reported in Fig. 7, where the number of

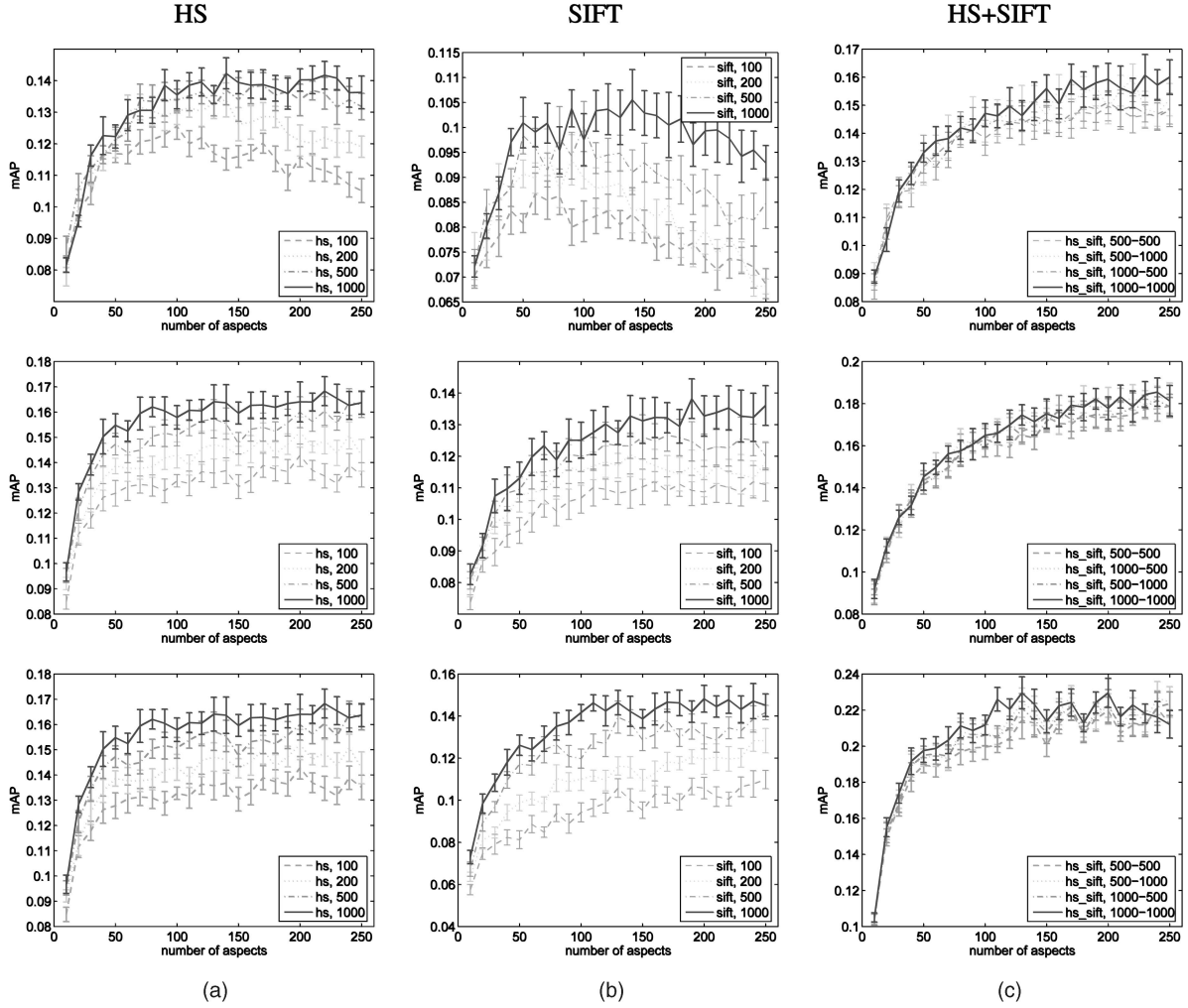

Fig. 7. Joint cross validation of the number of aspects and the number of K-means clusters for (a) the HS, (b) SIFT, and (c) HS+SIFT representations and for the three PLSA learning methods. The mAP values obtained for PLSA-MIXED (top row), PLSA-FEATURES (middle row), and PLSA-WORDS (bottom row) are given. The error bars show the standard deviation of the mAP values for 10 runs.

TABLE 3
mAP Values (in Percent) for the Six Methods When Combinations of HS and SIFT Features Are Used

	Blobs	HS	SIFT	HS+SIFT
propagation	7.8 (0.7)	9.0 (0.2)	9.4 (1.0)	13.1 (0.5)
CMRM [15]	11.5 (1.1)	10.7 (1.1)	7.9 (0.5)	13.4 (1.0)
SVD-COS [26]	12.9 (1.1)	12.9 (0.8)	10.7 (0.7)	16.6 (1.1)
PLSA-MIXED	5.8 (0.8)	10.2 (0.8)	7.5 (0.6)	11.9 (1.3)
PLSA-FEATURES	8.2 (0.7)	11.2 (1.0)	10.1 (0.8)	14.0 (1.3)
PLSA-WORDS	11.0 (0.9)	13.3 (1.0)	11.8 (1.1)	19.1 (1.2)

latent aspects is varied from 10 to 250 for the three PLSA-based approaches. The number of K-means clusters for quantizing the visual features is also varied and reported as a different line on each plot. The standard deviation over 10 cross-validation runs is shown with error bars.

The plots in Fig. 7 allow to decide the number of aspects and the number of clusters, given each PLSA learning methods and each image representation. The maximum number of K-means clusters seems to be a reasonable choice for the HS, SIFT, and HS+SIFT representations. The results reported in Section 6.4 are therefore computed with $N_h = 1,000$ and $N_s = 1,000$ clusters for the quantization of the HS and SIFT representations. Regarding the number of aspects N_z, the following values are chosen from the cross-validation experiments:

- PLSA-MIXED. $N_z = 140$ for HS, $N_z = 110$ for SIFT, and $N_z = 170$ for HS+SIFT.
- PLSA-FEATURES. $N_z = 170$ for HS, $N_z = 150$ for SIFT, and $N_z = 180$ for HS+SIFT.
- PLSA-WORDS. $N_z = 120$ for HS, $N_z = 110$ for SIFT, and $N_z = 120$ for HS+SIFT.

6.4 Overall Performance

The average of the mAP obtained from the 10 test sets with the hyperparameters estimated in Section 6.3 are shown in Table 3, where the performance of the Blob, HS, SIFT, and HS+SIFT representations for the six autoannotation methods presented in Sections 4.4 and 5 are reported. The standard deviation of the mAP over the 10 test sets is shown in parentheses. Note that the mAP values in Table 3 are consistently lower than the cross-validation values, because the retrieval tasks on which the mAP are now computed is more challenging: An average of 1,750 images for testing (versus an average of 520 images for cross validation) are ranked.

We see that the PLSA-MIXED approach particularly fails to produce an efficient probabilistic indexing of the test images for all the image representations. In particular, its performance is lower than the simple propagation baseline that relies on a direct image similarity computation. Using a concatenated representation of words and visual features, PLSA-MIXED attempts to simultaneously model the visual and textual modalities. It means that, intrinsically, PLSA-MIXED assumes that the two modalities have an equivalent importance in defining the latent space, which, as the results suggest, is not the most accurate assumption.

Except for the PLSA-MIXED case and the CMRM method when the SIFT representation is used, all methods achieve a higher performance than the propagation baseline. This shows that computing image similarity, although simple and intuitive, can only be considered as a low-quality choice for image annotation. It is, however, competitive with respect to the CMRM and PLSA-FEATURES methods, in particular, for the HS and HS+SIFT image representations.

All methods take advantage of the HS+SIFT combination: The performance of a single feature type is always lower than their combination, which confirms that HS and SIFT features encode complementary information. It is interesting to notice that the CMRM and SVD-COS methods achieve the best performance for the Blob representation, which is the representation they were originally evaluated on [26], [15]. These methods, however, do not produce the best overall performance, especially when compared to the PLSA-WORDS method. Furthermore, when the conditional probability distributions of the aspects, given the training documents d_i, $P(z|d_i)$, are learned from the visual features with PLSA-FEATURES, the estimation of the conditional distribution over words gives better results than PLSA-MIXED, as well as lower mAP values than the baseline methods.

Regarding PLSA-WORDS, our method achieves a lower mAP performance than the SVD-COS method for the Blob representation, but it exploits the HS, SIFT, and the HS+SIFT representations more efficiently than both the CMRM and the SVD-COS approaches. Furthermore, it consistently performs better than CMRM. The PLSA-WORDS model achieves the best mAP overall score when the concatenated SIFT and HS+SIFT representations are used. In the HS+SIFT case, the PLSA-WORDS improves over the SVD-COS method by 15 percent (relative improvement). This improvement in performance is statistically significant according to a paired-samples t-test with a p-value of 0.05, showing that the estimation of the aspect distribution based on the textual modality improves over both the linear-algebra-based SVD-COS method and the method that does not use aspect variables. We illustrate the word distributions estimated by the SVD-COS and PLSA-WORDS approaches in Fig. 8. For both images, the four words from the ground-truth annotation are in the top 20 words given by PLSA-WORDS, whereas only three of them are in the top 20 words given by SVD-COS. More importantly, the probability values are more contrasted in the PLSA-WORDS case. A few words are sharing a large proportion of the probability mass, which will be an advantage for ranking images based on these values. The SVD-COS approach estimates flatter word distributions, making the probabilities of a given word very similar across images. More annotation examples are given at http://www.idiap.ch/mam.

In the following sections, we analyze the performance of PLSA-WORDS, the best-performing model, in more detail.

6.5 Per-Word Performance

The histogram of the AP values obtained with PLSA-WORDS in Fig. 9b shows a marked difference in performance for different words: half of the words have an AP value higher than 0.14, 65 words have an AP value below 0.1, and 10 words have an AP value above 0.5. A similar trend can be observed with the baseline methods, as shown for the SVD-COS method in Fig. 9a. This important variation

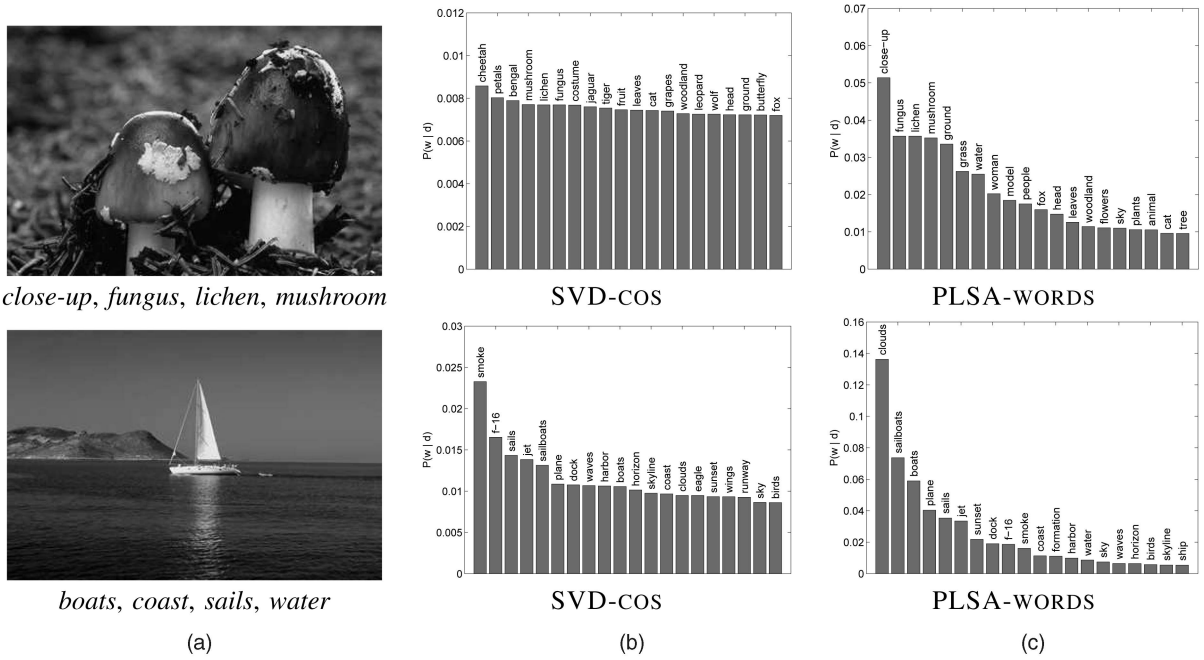

Fig. 8. The conditional probability distribution $P(w|d)$ inferred from two test images from their HS+SIFT representation with the SVD-COS and PLSA-WORDS approaches. The image and the ground-truth annotation are shown in (a). The top 20 words and their conditional probability are shown in (b) and (c) for the SVD-COS and PLSA-WORDS approaches, respectively.

goes unnoticed if only the mAP is reported, as done in part of the existing literature [17], [12].

The combined effect of three factors could explain why the system does not rank images satisfactorily for some words, whereas achieving a good performance for others. First, the number of training images per word ranges significantly in the data set from 21 (for *bay*, *candy*, *formula*, and so forth) to 1,124 (*water*), and obviously, the quality of a statistical model depends on the nature and the number of training examples. Second, all words have to be learned from the same set of visual features, which can be better suited for some concepts than for others. Third, the co-occurrence in text captions can have a combined influence with the two previous points; if a given word is correctly learned by the model because it is well represented by the visual features and has a sufficient number of training examples, other words that consistently co-occur with it could have a relatively high performance despite a low number of training examples. We investigate these three factors by analyzing individual word performance together with basic statistics from the training set.

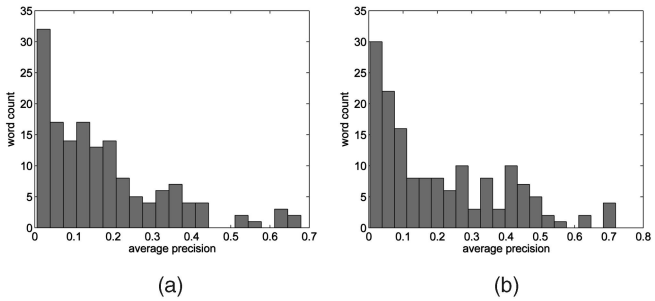

Fig. 9. Histogram of the 153 AP values for (a) SVD-COS and (b) PLSA-WORDS methods.

The number of training images and the AP for the 20 words with the best and the worst performance with the PLSA-WORDS model are shown in Fig. 10, which shows that there is a difference in the average number of examples for the 20 best performing words compared to the 20 worst performing words. The former have 106 training examples on the average, whereas the latter have an average of 29 examples. This fact suggests that the number of examples does indeed influence the performance of a word in general, because a low number of examples often do not allow comparison of the statistical variations of a word appearance.

However, we also see in Fig. 10 that, even though words have a comparable number of training examples, their respective performance is completely different. The words *polar*, *formula*, and *black* (Fig. 10a) have a high AP value (~ 0.5), whereas the words *river*, *woods*, and *road* (Fig. 10b) are part of the 20 words with the lowest AP (~ 0.015). The performance of a given word thus depends not only on the number of training examples, but also on the two other factors mentioned above.

In cases when the images a word is attached to depict a consistent visual content that is well represented by the feature set, the model can learn the representation from little training data. For instance, images that are annotated with the word *formula* contain distinctive visual features that can be captured from a relatively small number of examples (21 in the data set) while providing a high AP value of 0.5 for this word. Similarly, the word *polar* is mainly attached to winter images, which have a very distinctive white aspect, and is therefore well predicted (AP of 0.51) despite very few training examples (only 28). On the contrary, the words *reflection* and *museum* for instance are not correctly modeled because the corresponding image content can not be learned properly from 25 and 42 examples, respectively.

For models such as PLSA-WORDS that learn co-occurrences in image captions, there is a possibility to improve the

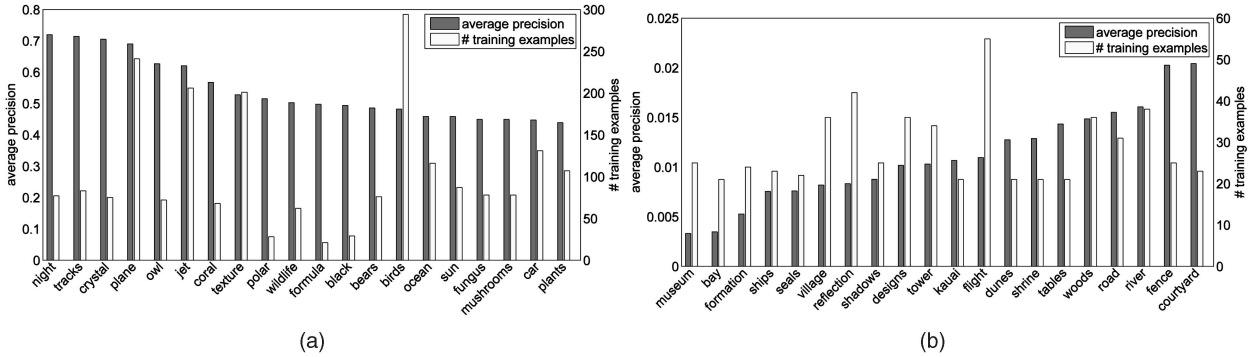

Fig. 10. AP and number of training examples for (a) the 20 best and (b) the 20 worst AP values (ranked in decreasing order in (a) and in increasing order in (b)) obtained with the PLSA-WORDS annotation.

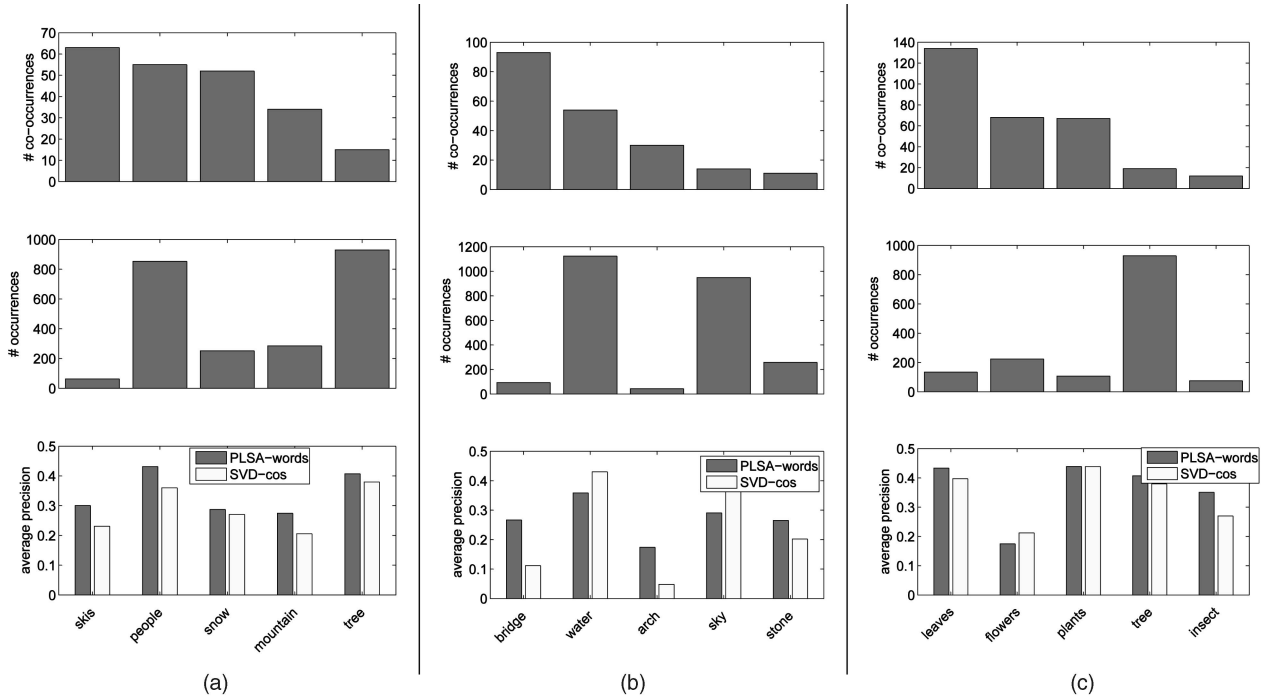

Fig. 11. Effect of word co-occurrences in captions for the words (a) *skis*, (b) *bridge*, and (c) *leaves*. The first row shows the number of times the four most frequently co-occurring words appear in the same caption as the word considered, the second row shows the total number of times each word appears in the training set, and the third row shows the AP of these words for the PLSA-WORDS (dark gray) and the SVD-COS (white) models.

prediction of infrequent words from their co-occurring words. We show three examples of this effect in Fig. 11 for the words *skis*, *bridge*, and *leaves*. For these three words, the four words that co-occur the most with each of them are reported, as well as different statistics, including the number of times they co-occur with the word considered (top row), the number of times they appear in the training set (middle row), and their respective AP values (bottom row). Regarding the first example, although the word *skis* is only represented by 63 examples, the fact that it co-occurs quite often with more frequent words like *people* (which appears in 853 examples), *snow* (252 examples), and *mountain* (82 examples), allows PLSA-WORDS to predict *skis* with a high AP (0.28). For the second example, the word *bridge* only has 93 examples but is well predicted by the PLSA-WORDS model, because it co-occurs with words that have more examples in the training set, like *water* (which occurs in 1,124 examples), *sky* (949 examples), and *stone* (258 examples). For the last example, the word *leaves* is predicted with an AP of 0.43 by PLSA-WORDS, although there are only 134 *leaves* image examples. The fact that the word *leaves* co-occurs quite frequently with the words *flowers* (appearing in 224 examples) and *tree* (929 examples) also illustrates why a model that captures co-occurrence information at the caption level performs better than a model that does not model this information explicitly. In the last two examples, SVD-COS fails to take advantage of the co-occurrence with more frequent words, as PLSA-WORDS does.

6.6 Combination of Features

To observe in more detail the benefit of combining HS and SIFT features for PLSA-WORDS, their individual and combined effects on the AP of 10 representative words is shown in Fig. 12. These 10 words are selected to illustrate the different interesting behaviors that are observed when SIFT (dark gray), HS (gray), or both (white) are used.

As a general trend, we see that words that are rather well defined by color regions have higher AP values when the

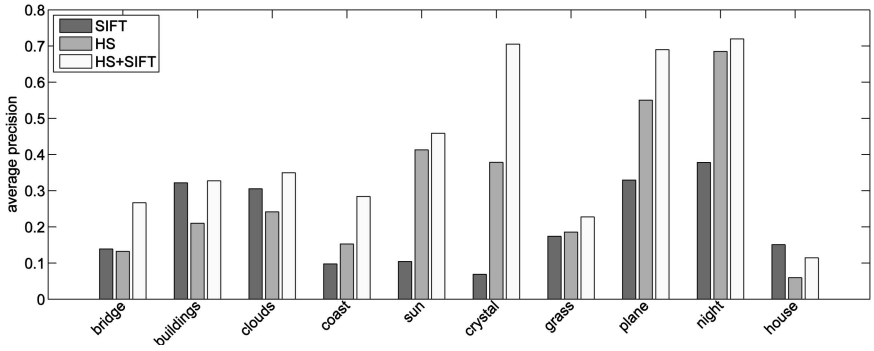

Fig. 12. AP of 10 selected words when SIFT (*dark gray*), HS (*gray*), and HS+SIFT (*white*) features represent the image. Depending on the word, the AP values are higher for one of the two representations and the combination of both improves in general.

HS representation is used, compared to SIFT local patches. In Fig. 12, images annotated with words such as *sun*, *crystal*, *plane*, and *night* depict colored regions and are therefore well represented by the HS features. As shown in the Fig. 12, the AP of these words for a retrieval system based on the HS representation outperforms the same system based on the SIFT representation. This is a somewhat expected result. For instance, images annotated by the word *sun* present rather nondistinctive image structures but contain very specific colors. Similarly, *crystal* images have a large variety of textures but present distinctive colors. The AP of this word is therefore higher when HS features are used. The word *plane* also happens to be better represented by HS features as shown in Fig. 12, which could be at first glance counterintuitive. However, the word *plane* consistently appears in the context of blue sky, which is well identified by the HS representation.

On the contrary, if a word corresponds to images that contain specific textures, the SIFT representation becomes more informative and results in better image ranking. This can be observed in Fig. 12, where the AP values for the words *buildings*, *clouds*, and *house* are higher when the SIFT (instead of the HS) representation is used. All these images contain structures that are less well represented by HS elements, which encode color information. Based on local gray-scale edge directions, the SIFT patches can efficiently depict parts of these structures and allow to discriminate between, for example, a white house and a polar bear that would be represented by a similar HS histogram. In Fig. 12, we see that the *house* AP values are more than two times bigger for the SIFT than for the Blob representation.

As already shown in Table 3, the concatenation of the HS and SIFT representations provides the best performance. More precisely, it improves the AP of 121 words compared to the SIFT-only representation and 121 words compared to the HS-only representation. This complementarity can be analyzed in more detail on the 10 words considered in Fig. 12. The concatenation of HS and SIFT features improves the AP of nine of the 10 words in Fig. 12 and on all of them on the average, as shown in Table 3.

Regarding limitations of the HS+SIFT combination, note that, for some words like *house* in Fig. 12, combining the SIFT and the HS representations actually produces a worse image ranking than the SIFT-only case. This indicates that some ambiguity is introduced by the HS features in the related images, making them more similar to other images that are annotated with different words. Better mechanisms for data fusion could thus potentially improve the system performance, because a few words are better represented by one of the two feature types than by their simple concatenation. The fact that one model is learned for all the words does not allow a basic word-dependent weighting of the features and more elaborate schemes have to be explored in the future.

7 Conclusion

We presented three alternative algorithms to learn a PLSA model for annotated images and evaluated their ability for cross-media image indexing. The learning methods differ in which of the textual or the visual modality is dominant to learn the mixture of aspects for an image and its text caption, and these differences influence the accuracy of the inferred semantic indexes. The best retrieval performance is achieved when the mixture of latent aspects is learned from the text captions (our PLSA-WORDS model), creating semantically meaningful aspects.

We also proposed to combine quantized local color information with quantized local image descriptors, demonstrating their complementarity and their improved performance when compared to the standard Blob representation. The performance of all the models was improved by the use of this combined image representation, which depicts an image as a set of local-color-based regions and local-texture-based regions. In particular, the PLSA-WORDS model achieved the best performance with respect to recent methods.

The quality of the image ranking greatly varies depending on the query, and we analyzed the possible factors in the case of the PLSA-WORDS model. Besides the difference in the number of training examples or the suitability of the visual features to represent a given concept, we have shown strong indications that PLSA-WORDS can take advantage of the co-occurrence of words in the text captions of the training images.

Acknowledgments

This work was funded by the Swiss National Science Foundation through the MULTI project and the Swiss National Center of Competence in Research on Interactive Multimodal Information Management (NCCR (IM)2). The authors thank Pedro Quelhas and Jean-Marc Odobez for discussions.

REFERENCES

[1] S. Agarwal, A. Awan, and D. Roth, "Learning to Detect Objects in Images via a Sparse, Part-Based Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, pp. 1475-1490, 2004.

[2] K. Barnard, P. Duygulu, N. Freitas, D. Forsyth, D. Blei, and M.I. Jordan, "Matching Words and Pictures," *J. Machine Learning Research,* vol. 3, pp. 1107-1135, 2003.

[3] D. Blei and M. Jordan, "Modeling Annotated Data," *Proc. Int'l Conf. Research and Development in Information Retrieval,* Aug. 2003.

[4] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

[5] T. Brants, "Test Data Likelihood for PLSA Models," *Information Retrieval,* vol. 8, pp. 181-196, 2005.

[6] W. Buntine, "Variational Extensions to EM and Multinomial PCA," *Proc. European Conf. Machine Learning,* Aug. 2002.

[7] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, pp. 1026-1038, 2002.

[8] E.Y. Chang, K. Goh, G. Sychay, and G. Wu, "CBSA: Content-Based Soft Annotation for Multimodal Image Retrieval Using Bayes Point Machines," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 13, pp. 26-38, 2003.

[9] P. Duygulu, K. Barnard, J.F.G. de Freitas, and D.A. Forsyth, "Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary," *Proc. European Conf. Computer Vision,* May 2002.

[10] J. Dy, C. Brodley, A.C. Kak, L. Broderick, and A. Aisen, "Unsupervised Feature Selection Applied to Content-Based Retrieval of Lung Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, pp. 373-378, 2003.

[11] L. Fei-Fei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* June 2005.

[12] S.L. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli Relevance Models for Image and Video Annotation," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* June 2004.

[13] A. Girgensohn, J. Adcock, and L. Wilcox, "Leveraging Face Recognition Technology to Find and Organize Photos," *Proc. ACM SIGMM Int'l Workshop Multimedia Information Retrieval,* 2004.

[14] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning,* vol. 42, pp. 177-196, 2001.

[15] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic Image Annotation and Retrieval Using Cross-Media Relevance Models," *Proc. 26th Int'l Conf. Research and Development in Information Retrieval,* Aug. 2003.

[16] J. Jeon and R. Manmatha, "Using Maximum Entropy for Automatic Image Annotation," *Proc. IEEE Int'l Conf. Image and Video Retrieval,* July 2004.

[17] V. Lavrenko, R. Manmatha, and J. Jeon, "A Model for Learning the Semantics of Pictures," *Proc. Ann. Conf. Neural Information Processing Systems,* Dec. 2003.

[18] J. Li and J.Z. Wang, "Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, pp. 1075-1088, 2003.

[19] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision,* vol. 60, pp. 91-110, 2004.

[20] F. Monay and D. Gatica-Perez, "On Image Auto-Annotation with Latent Space Models," *Proc. ACM Int'l Conf. Multimedia,* Nov. 2003.

[21] F. Monay, P. Quelhas, D. Gatica-Perez, and J.-M. Odobez, "Constructing Vision Models with a Latent Space Approach," *Proc. PASCAL Workshop Subspace, Latent Structure and Feature Selection Techniques: Statistical and Optimisation Perspectives,* Feb. 2005.

[22] Y. Mori, H. Takahashi, and R. Oka, "Image-to-Word Transformation Based on Dividing and Vector Quantizing Images with Words," *Proc. Int'l Workshop Multimedia Intelligent Storage and Retrieval Management,* Oct. 1999.

[23] H. Mueller, S. Marchand-Maillet, and T. Pun, "The Truth about Corel—Evaluation in Image Retrieval," *Proc. Int'l Conf. Image and Video Retrieval,* July 2002.

[24] W. Niblack, R. Barber, W. Equitz, M. Flicker, E. Glasman, D. Petkovic, P. Yanker, and C. Faloutsos, "The QBIC Project: Query Images by Content Using Color, Texture and Shape," *Proc. SPIE Conf. Storage and Retrieval for Image and Video Databases,* Feb. 1993.

[25] M. Ortega, Y. Rui, K. Chakrabarti, S. Mehrotra, and T.S. Huang, "Supporting Similarity Queries in MARS," *Proc. ACM Int'l Conf. Multimedia,* Nov. 1997.

[26] J.-Y. Pan, H.-J. Yang, P. Duygulu, and C. Faloutsos, "Automatic Image Captioning," *Proc. IEEE Int'l Conf. Multimedia and Expo,* June 2004.

[27] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L.V. Gool, "Modeling Scenes with Local Descriptors and Latent Aspects," *Proc. IEEE Int'l Conf. Computer Vision,* Oct. 2005.

[28] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* June 1997.

[29] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman, "Discovering Object Categories in Image Collections," *Proc. IEEE Int'l Conf. Computer Vision,* Oct. 2005.

[30] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval: The End of the Early Years," *IEEE Trans. Pattern Analysis Machine Intelligence,* vol. 22, pp. 1349-1380, 2000.

[31] J. Smith, C. Lin, M. Naphade, A. Natsev, and B. Tseng, "Multimedia Semantic Indexing Using Model Vectors," *Proc. IEEE Int'l Conf. Multimedia and Expo,* July 2003.

[32] J.R. Smith and S.-F. Chang, "Visualseek: A Fully Automated Content-Based Image Query System," *Proc. ACM Int'l Conf. Multimedia,* Nov. 1996.

[33] K. Thieu and P. Viola, "Boosting Image Retrieval," *Int'l J. Computer Vision,* vol. 56, pp. 17-36, 2004.

[34] P. Viola and M. Jones, "Robust Real-Time Face Detection," *Int'l J. Computer Vision,* vol. 57, pp. 137-154, 2004.

[35] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* Dec. 2001.

Florent Monay received the MS degree in microengineering in 2002 from the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. He then joined the IDIAP Research Institute in Martigny, Switzerland, as a research assistant and obtained his PhD from the Ecole Polytechnique Fédérale de Lausanne (EPFL) in 2007. He is now working at the Signal Processing Institute (EPFL) as a research scientist. His interests cover multimedia information retrieval, text retrieval, computer vision, and statistical machine learning.

Daniel Gatica-Perez received the BS degree in electronic engineering from the University of Puebla, Mexico, in 1993, the MS degree in electrical engineering from the National University of Mexico in 1996, and the PhD degree in electrical engineering from the University of Washington, Seattle, in 2001. He joined the IDIAP Research Institute in 2002, where he is now a senior researcher. His interests include multimedia signal processing and information retrieval, computer vision, and statistical machine learning applied to these domains. He currently is an associate editor of the *IEEE Transactions on Multimedia*. He is a member of the IEEE.