IDIAP RESEARCH REPORT

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

Florent Monay [a] Daniel Gatica-Perez [a]

IDIAP–RR 04-30

June 2004

submitted for publication

[a] IDIAP, Martigny, Switzerland

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP Research Report 04-30

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

Florent Monay Daniel Gatica-Perez

June 2004

submitted for publication

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

1 Introduction

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

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

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

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

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

2 Data Representation

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

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

Image. We use two common image representations.

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

Blobs [3] : The normalized cut segmentation algorithm is applied to the image set, and the resulting regions are represented by color, texture, shape, size, and position descriptors. The K-means clustering algorithm is applied to all the computed descriptors, quantizing the image regions into a B-dimensional *blob vector-space* (B=500, same notation as RGB).

3 The PLSA model

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

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

$$P(x_j, d_i) = P(d_i) \sum_k^K P(z_k \mid d_i)P(x_j \mid z_k). \tag{1}$$

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

On the contrary, the K-by-M table $P(z \mid d)$ is only relative to the M training documents. $P(z_k \mid d_i)$ is the probability that a particular document d_i contains the aspect z_k, which does not carry any a priori information about the probability of aspect z_k being expressed in any unseen document.

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

$$\mathcal{L} = \prod_i^M \prod_j^N P(d_i) \sum_k^K P(z_k \mid d_i)P(x_j \mid z_k)^{n(d_i, x_j)}, \tag{2}$$

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

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

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

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

4 PLSA-based Annotation

PLSA has been recently proposed as a model for image auto-annotation [6]. Referred here as PLSA-MIXED, it showed (somewhat surprisingly) a poor annotation performance with respect to basic non

probabilistic methods [6]. We propose here a new application of PLSA to automatic image annotation, motivating our approach by an analysis of the limitations of PLSA-MIXED.

4.1 PLSA-mixed

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

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

4.2 Problems with PLSA-mixed

Using a concatenated representation, PLSA-MIXED attempts to simultaneously model the visual and textual modalities, intrinsically assuming that the two modalities have an equivalent importance in defining the latent space. This has traditionally been the assumption in most previous works [4]. However, an analysis of the captions and the image features in the Corel dataset (described in Section 5) emphasizes the difference between the keywords and the visual features occurrences. Figure 1 shows two pair-wise similarity matrices for a set of annotated images ordered by topics, as defined by the human-generated CD organization of the Corel collection. The matrices represent the cosine similarity between each document in the keyword space (left), and the visual feature space (right). We can see that the keyword matrix has a sharp block-diagonal structure, each corresponding to a consistent cluster of images, while the visual feature matrix depicts a much less clear structure.

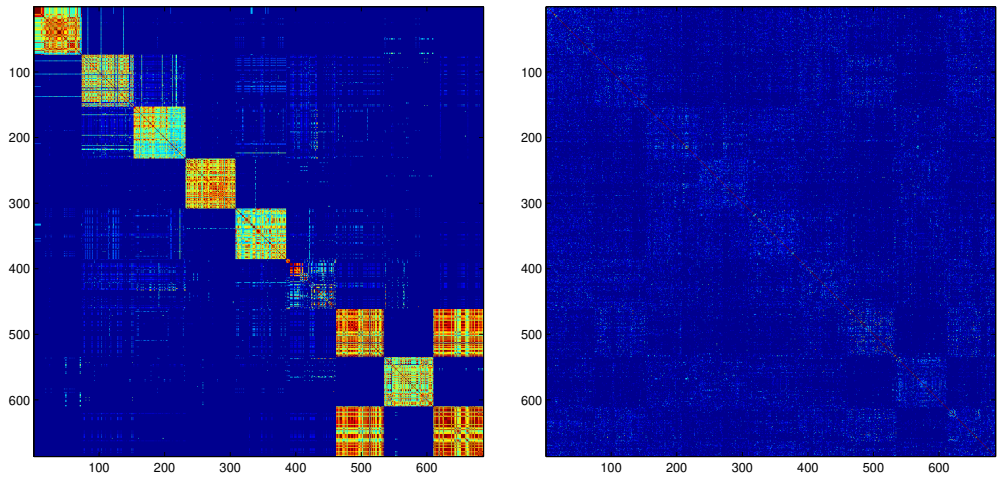

Figure 1: Pair-wise similarity matrices for a set of manually ordered documents (nine CDs from Corel). The left matrix is the textual modality, the right matrix is the visual modality (blobs features).

Of course, Figure 1 does not prove that no latent representation exists for the visual features, but it clearly suggests that in general, two PLSA separately applied on each modality would define

two distinct latent representations of the same document. For example, color co-occurrence happens across images, but does not necessarily mean that the corresponding images are semantically related. If PLSA-MIXED relies too strongly on visual features to learn the latent aspects, the auto-annotation model might end up predicting almost random keywords if these aspects have high probabilities given the image to annotate. Moreover, assuming that no particular importance is given to any modality, the amount of visual and textual information needs to be balanced in the concatenated representation of an annotated image. This restricts the size of the visual representation, as the number of keywords per image is usually limited (e.g., an average of 3 for the Corel data). A typical aspect from PLSA-MIXED where images are relatively consistent in terms of visual features (dominant colors: green, red, yellow, black, gray) but not semantically is shown in Figure 2.

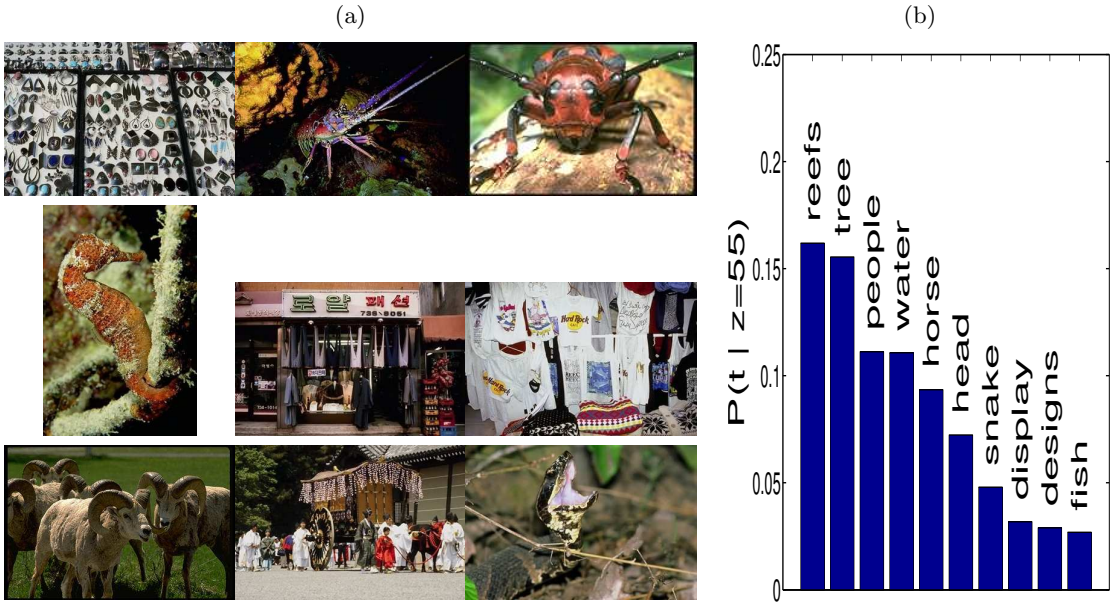

Figure 2: One semantically meaningless aspect from PLSA-MIXED: the 9 most probable images in the training set, and the 10 most probable keywords with their corresponding probability $P(t \mid z)$.

4.3 Our approach: PLSA-words

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

Parameter learning

1. A first PLSA model is completely trained on the set of image captions to learn both $P(t \mid z)$ and $P(z \mid d)$. Figure 3 illustrates two aspects automatically learned on the textual modality, with their most probable training images (a and c), and their corresponding distribution over keywords $P(t \mid z)$ (b and d). These examples show that this first PLSA can capture semantically meaningful aspects from the data.

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

Annotation by inference

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

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

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

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

Our method differs from existing ones, including Gaussian-Multinomial LDA (GM-LDA) [2], which uses two distinct latent spaces for keywords and image features, and from Correspondence LDA (Corr-LDA)[2], which although assumes only one latent space, conditions the keywords during training on underlying factors that generated the image features, in a process somewhat reverse in spirit to our approach.

5 Performance evaluation

5.1 Data

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

5.2 Performance measures

No commonly agreed image auto-annotation measure exists. We evaluated our method with three different measures, but restrict the discussion to the two measures described below for space reasons. Precision/recall results, which confirm what the other measures suggest, are available on a website (ommitted for double-blind review).

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

Normalized Score (NS). With r and n defined as above, the normalized score is defined as $NS = r/n - (p - r)/(N - n)$, where N is the vocabulary size and p is the number of predicted keywords [4]. The average NS is computed over a set of images for a varying number of predicted keywords. Its maximum (the higher the better), and the value at which it occurs (in number of predicted words, the lower the better) are reported here.

5.3 Results and Discussion

We compare the two PLSA-based methods described in Section 4.1 and 4.3, and three other methods : EMPIRICAL, LSA and PLSA-SPLIT. EMPIRICAL simply uses the empirical keyword distribution from the training set to predict the same set of keywords regardless of the image content. LSA (i.e., non-probabilistic LSA) was reported in the literature as better than PLSA-MIXED in NS terms [6]. PLSA-SPLIT is the *unlinked* equivalent of PLSA-WORDS, for which two distinct sets of parameters $P_t(z \mid d)$ and $P_v(z \mid d)$ are learned for each modality. A value of $K = 100$ for the latent space

dimensionality was used for all the reported results (except EMPIRICAL which does not need it). The average AA and the maximum NS are presented in Tables 1 and 2, respectively. All results are averaged over the 10 image subsets.

Method	BLOBS	RGB
EMPIRICAL	0.191 (0.012)	0.191 (0.012)
LSA	0.140 (0.009)	0.178 (0.009)
PLSA-SPLIT	0.113 (0.017)	0.121 (0.019)
PLSA-MIXED	0.221 (0.011)	0.217 (0.024)
PLSA-WORDS	0.279 (0.014)	0.276 (0.014)

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

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

We confirmed that, as reported by [6], the PLSA-MIXED maximum NS is lower than the one obtained with LSA. On the other hand, the evaluation with AA, which measures the quality of a much smaller but more realistic annotation (3.1 words in average), ranked PLSA-MIXED as better than LSA.

The ranking of the three PLSA-based methods emphasizes the importance of a well defined link between textual and visual modalities. PLSA-SPLIT naively assumes no link between captions and images and models them separately. No relation between the two latent space definitions exists, which explains why PLSA-SPLIT performs worse than the simplest EMPIRICAL method. The PLSA-MIXED method certainly introduces a degree of interaction between text and image by concatenating the two modalities. This connexion translates into a significant improvement over PLSA-SPLIT in terms of both AA and NS.

PLSA-WORDS outperforms both PLSA-SPLIT and PLSA-MIXED. PLSA-WORDS makes an explicit link between visual features and keywords, learning the latent aspect distribution in the keyword space, and fixing these parameters to learn the distribution of visual features. This results in the definition of more semantically meaningful aspects, and drives the model to predict more consistent keywords. Overall, our method performs consistently better than all the other methods for all the measures (AA, NS, and precision/recall, not shown). The relative AA improvement for the blobs features is 99% w.r.t. LSA, and 26% w.r.t. PLSA-MIXED (respectively 55% and 21% for the RGB features). In Table 2, also note that the maximum NS is attained at a considerably lower number of predicted words for our approach.

Method	BLOBS	RGB
EMPIRICAL	0.429 (0.016) [36.2]	0.429 (0.016) [35.6]
LSA	0.521 (0.013) [40.6]	0.539 (0.011) [38.7]
PLSA-SPLIT	0.273 (0.020) [43.8]	0.281 (0.016) [34.3]
PLSA-MIXED	0.463 (0.018) [37.2]	0.475 (0.020) [34.3]
PLSA-WORDS	0.549 (0.013) [28.8]	0.556 (0.011) [30.6]

Table 2: Average maximum NS over the 10 subsets. The variance is given in parentheses. The average number of predicted words corresponding to the maximum NS is in brackets.

6 Conclusion

We proposed a new PLSA-based image auto-annotation method, which uses two linked PLSA models to represent the textual and visual modalites of an annotated image. This allows a different processing of each modality while learning the parameters, and makes the definition of a more semantically meaningful latent space possible. We compared our method to previously proposed models using various performance measures, and showed that the new model significantly improves previous latent space methods based on a concatenated textual+visual representation. The benefits of the model, for a larger keyword vocabulary size and more complex visual features, has to be investigated.

Acknowledgments

This work was funded by the swiss NCCR on Interactive Multimodal Information Management (IM)2. The images used in this study belong to the Corel stock photo collection ©.

References

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

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

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

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

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

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

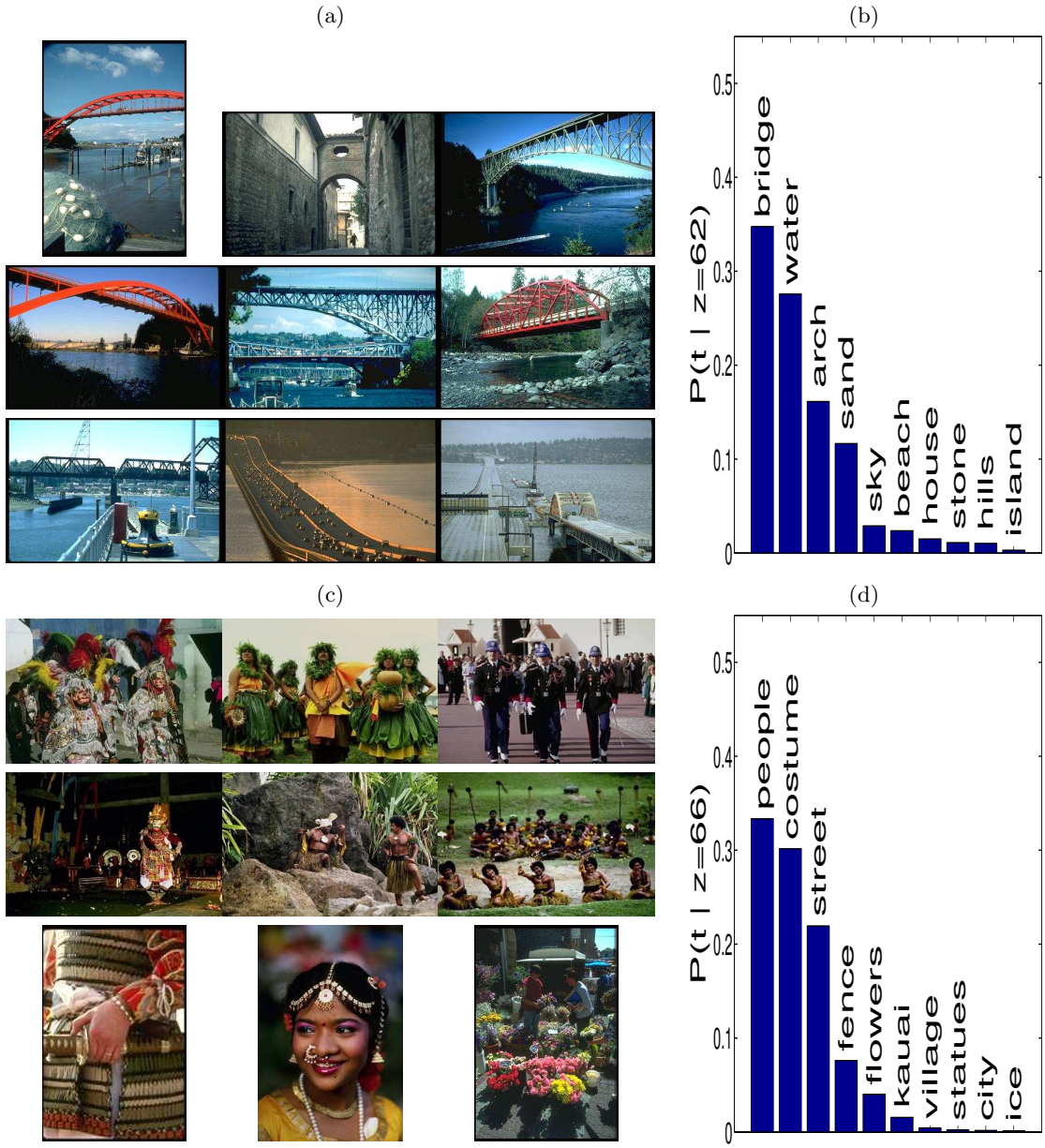

Figure 3: Two different aspects from PLSA learned on words: the 9 most probable images in the training set (from $P(z \mid d)$), and the 10 most probable keywords with their corresponding probability $P(t \mid z)$.