Geo-located image analysis using latent representations

M. Cristani A. Perina U. Castellani
V. Murino
Dipartimento di Informatica, Università di Verona
37134, Verona, Italy
marco.cristani|alessandro.perina|umberto.castellani|vittorio.murino|@univr.it

Abstract

Image categorization is undoubtedly one of the most challenging open problems faced in Computer Vision, far from being solved by employing pure visual cues. Recently, additional textual "tags" can be associated to images, enriching their semantic interpretation beyond the pure visual aspect, and helping to bridge the so-called semantic gap. One of the latest class of tags consists in geo-location data, containing information about the geographical site where an image has been captured. Such data motivate, if not require, novel strategies to categorize images, and pose new problems to focus on. In this paper, we present a statistical method for geo-located image categorization, in which categories are formed by clustering geographically proximal images with similar visual appearance. The proposed strategy permits also to deal with the geo-recognition problem, i.e., to infer the geographical area depicted by images with no available location information. The method lies in the wide literature on statistical latent representations, in particular, the probabilistic Latent Semantic Analysis (pLSA) paradigm has been extended, introducing a latent aspect which characterizes peculiar visual features of different geographical zones. Experiments on categorization and geo-recognition have been carried out employing a well-known geographical image repository: results are actually very promising, opening new interesting challenges and applications in this research field.

1. Introduction

Categorizing pictures in an automatic and meaningful way is the key challenge in all the retrieval-by-content systems [19]. Early categorization techniques used uniquely visual cues to build compact images descriptions, with the aim of enhancing those image elements relevant for the categorization while disregarding the others. Recently, this classical framework has been improved with the use of text labels or *tags*, associated to the images [1], usually provided by a human user, devoted to constrain the number of ways an automatic system can categorize an image.

This structure has been further updated with the introduction on the market of several cheap Global Positioning System (GPS) devices mounted on cameras. Such devices automatically assign tags to the captured pictures, indicating the geographical position (latitude, longitude) of the shot. This capability leaded to the creation of successful global repositories for *geo-located* images, such as Panoramio[1], which now maintains huge amounts of (mainly outdoor) pictures.

This situation encourages the design of novel strategies to categorize images, and pose new problems to focus on. For example, *geo-categorization* algorithms can be designed, taking into account the geographic location of the images, other than their visual aspect. The underlying idea is to individuate particular regions, here called *geo*-categories, in which are located visually similar pictures. In this way, the content of a geo-located image database can be visualized by means of few representative images per geo-category. Beyond the mere visualization, the geo-categorization gives also insight on the content of the image repository: being the pictures personal observations of the environment, we have that each geo-category encodes visual aspects of the territory relevant for the community.

Another interesting and hard problem to deal with in this context is the *geo-recognition* of images, where the goal is to infer the geographical area in which a non geo-tagged picture has been acquired. This task is useful in different fields: in the context of web content mining, where the extraction of geographical location information from a web page has recently become an important task [20]. Geo-recognition can also be useful in the forensic area, for instance, to constrain the possible zones in which a picture has been taken.

An issue similar to the geo-recognition was faced few years ago, under the name of *location recognition task*, as

[1] http://www.panoramio.com

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

an open research contest [2]. The geo-recognition task was in this case faced by taking into account 3D reconstruction methods, due to the fact that the input images (taken by a calibrated camera) represented urban scenes with overlapped fields of view.

In our situation, the task is much harder, never faced before without resorting to additional textual data: here we deal with pictures portraying heterogeneous scenes, captured under completely different and unknown acquisition conditions (e.g., different poses, different time of the day and weather). Therefore, it is reasonable to drop relying on the geometric content encoded in the pictures, and to build a recognition technique based on the 2D image pictorial features.

In this paper, we propose a novel statistical framework aimed at the geo-located image categorization and geo-recognition, based on *latent* representations. Latent or topic models, such as the ones built by the probabilistic Latent Semantic Analysis (pLSA)[7] or by the Latent Dirichlet Allocation (LDA) [2], were originally used in the text understanding community for unsupervised topic discovery in a corpus of documents. In Computer Vision, topic models have been used to discover scene classes, or visual topics, from a collection of unlabeled images. Here, the input data are the 'bags of visterms', i.e., histograms of local visual aspects extracted from the images. Co-occurrences of visterms in the images form the topics, which can be considered as higher level descriptions of the images. As a result, images can be categorized according to the topics they contain.

The work takes inspiration by the pLSA framework, extending it by introducing a *geo-topic* in addition to the (visual) topic. The geo-topic is associated to the visual topic information via a conditional dependency relation: in practice, each geo-topic describes visually a gaussian-shaped geographical area by means of a distribution on the visual topics. We call this paradigm *Location Dependent-pLSA* (LD-pLSA).

In the following, we will detail how the LD-pLSA performs when applied to a consistent database, built from the Panoramio repository, also providing comparative tests of categorization. Therefore, we will show geo-recognition results comparing our method with other ad-hoc approaches.

The rest of the paper is organized as follows. In Sec. 2, notation and background notions on pLSA are reported. In Sec. 3, the LD-pLSA framework is fully detailed, showing how geo-located image categorization and geo-recognition can be formulated. Sec.4 illustrates the experiments carried out to validate the proposed approach. Finally, in Sec.5, conclusions and future perspectives are drawn.

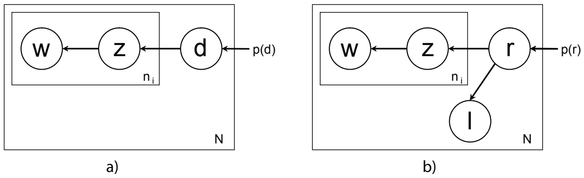

Figure 1. Graphical models: a) pLSA; b) LD-pLSA

2. Probabilistic Latent Semantic Analysis

Originally suited for analyzing text corpi, pLSA [7] has been extensively investigated in the modeling of image collections. The input is a dataset of N images $\{d_i\}, i = 1,..., N$, each described as an histogram of local regions found by interest operators, whose appearance has been quantized into M visual words $w_j, j = 1,..., M$. The dataset is thus summarized by a co-occurrence matrix of size $M \times N$, where the entry $<w_j, d_i>$ indicates the number of occurrences of the visual word w_j in the document d_i, also addressed as $n(w_j, d_i)$. Each document d_i has n_i words. The presence of a word w_j in the document d_i is mediated by a latent *topic* variable, $z \in Z = \{z_1,..., z_Z\}$, also called *aspect* class, *i.e.*,

$$P(w_j, d_i) = \sum_{k=1}^{Z} P(w_j|z_k)P(z_k|d_i)P(d_i). \qquad (1)$$

In practice, the topic z_k is a probabilistic co-occurrence of words encoded by the distribution $P(w|z_k)$, $w = \{w_1,..., w_M\}$, and each image d_i is compactly (usually, $Z < M$) modeled as a probability distribution over the topics, *i.e.*, $P(z|d_i)$, $z = \{z_1,..., z_Z\}$; $P(d_i)$ accounts for varying number of words in the images, *i.e.* $P(d_i) = n_i / \sum_{i=1}^{N} n_i$. The graphical model of pLSA is shown in Fig.1a. The hidden distributions of the model, $P(w|z)$ and $P(z|d)$, are learnt using Expectation-Maximization (EM) [5], maximizing the model data-likelihood L:

$$L = \prod_{i=1}^{N} \prod_{j=1}^{M} P(w_j, d_i)^{n(w_j, d_i)} \qquad (2)$$

The E-step computes the posterior over the topics, $P(z|w, d)$, and the M-step updates the hidden distributions. Once the model has been learnt, the most used inference, also called recognition inference, estimates the topic distribution of a novel image. Here, the learning algorithm is applied fixing the previously learnt distribution $P(w|z)$ and estimating $P(z|d)$ for the query image. For a deeper review of pLSA, see [7].

Different pLSA extensions are present in literature: the most known is the Latent Dirichlet Allocation (LDA) [2], which adds a sparse Dirichlet prior for the topic probabilities $P(z|d_i)$. LDA is computationally more demanding than

[2] *Where Am I?* ICCV Computer Vision Contest, please see `http://research.microsoft.com/iccv2005/Contest/`

pLSA, with a comparable accuracy. The benefit of LDA emerges when few images and several topics are considered, but this is not our case, as we will see in the following.

3. The proposed method: LD-pLSA

In LD-pLSA, we have N geo-located images $\{l_i\}$,i = 1,..., N, indexed as couples of latitude and longitude coordinate values. Associated to each i-th couple, we have a $M \times 1$ counting array $n(w, l_i)$ of visual words, for a total of n_i words. Our purpose is to simultaneously extract *two* different kinds of latent classes underlying the observed data: the *visual* topic and the *geo*-topic classes. The visual topic class z encodes, as in the original pLSA framework, probabilistic co-occurrences of visual words. The geo-topic class $r \in R = \{r_1,..., r_R\}$ serves to partition the entire geographic area spanned by the geo-located images into regions, each one characterized by a specific set of visual topics. The joint distribution over visual words and geo-located images can be factorized as follows:

$$P(w_j,l_i) = \sum_{k=1}^{Z} P(w_j|z_k) \sum_{c=1}^{R} P(z_k|r_c) p(r_c|l_i) P(l_i) \quad (3)$$

In the formula above, we have Z and R visual and geo-topic instances, respectively. The meaning of $P(w|z)$ is the same as in the original pLSA; the distribution $P(z|r)$ is a $Z \times R$ matrix, where each entry $<z_k, r_c>$ represents the probability that visual topic z_k is present in the region r_c. The density $p(r|l)$ is encoded as a $R \times N$ matrix, and models the likelihood of being in a particular region r_c, given the geo-located image l_i. Finally, $P(l)$ functions as the classical document-distribution $P(d)$ of pLSA.

An alternative consistent factorization of the model can be obtained by applying the Bayes theorem to $p(r|l)p(l)$, as done in a similar way in the classical pLSA in [7], obtaining the joint probability

$$P(w_j, l_i) = \sum_{k=1}^{Z} P(w_j|z_k) \sum_{c=1}^{R} P(z_k|r_c) p(l_i|r_c) p(r_c) \quad (4)$$

which permits to characterize geo-topics in both visual sense by means of $P(z_k|r_c)$, and under a topological aspect through $p(l_i|r_c)$. The correspondent graphical model is depicted in Fig.1b. In this paper, we assume a Gaussian form for $p(l|r)$, *i.e.*,

$$p(l_i|r_c) = \mathcal{N}(l_i; \mu_{r_c}, \Sigma_{r_c}) \quad (5)$$

where the parameters μ_{r_c}, Σ_{r_c} indicate the mean location of the r_c-th region and the associated spread, respectively. Finally, $p(r_c)$ is a discrete distribution over the region variable.

As compared to pLSA, the biggest difference is that LD-pLSA introduces a conditional independence relation between the visual topic z_k and the geo-located image l_i, given the geo-topic value r_c. This means that here the visual topic is evaluated as a global characteristic that influences *all* the images that lie within a region.

3.1. Model fitting with the EM algorithm

Given a set of training data, our model is learned by maximizing the data log-likelihood

$$LL = \sum_{i=1}^{N} \sum_{j=1}^{M} n(w_j, l_i) \log P(w_j, l_i) \quad (6)$$

where the joint probability is factorized as in Eq.4. The learning equations can be straightforwardly derived from those of pLSA, permitting a fast training via exact EM. In the E-step, the Bayes formula is applied in the parameterization of Eq.4, obtaining the posterior[3]

$$P(z, r|w, l) = \frac{P(w|z)P(z|r)p(l|r)p(r)}{\sum_{z,r} P(w|z)P(z|r)p(l|r)p(r)} \quad (7)$$

In the M-step, the expected complete data likelihood has to be maximized, which is

$$\begin{aligned} E[L] &= \sum_{w,d} n(w, d) \sum_{z,r} P(z, r|w, l) \\ &\cdot [\log P(w|z)P(z|r)p(l|r)p(r)] \quad (8) \end{aligned}$$

The maximization of $E[L]$ can be derived straightforwardly for the parameters describing $P(w_j|z_k)$, $P(z_k|r_c)$, $P(r_c)$ and $p(l_i|r_c)$, employing Lagrange multipliers where necessary. The M-step re-estimation equations are thus:

$$P(w|z) = \frac{\sum_l n(w, l) \sum_r P(z, r|w, l)}{\sum_{w,l} n(w, l) \sum_r P(z, r|w, l)} \quad (9)$$

$$P(z|r) = \frac{\sum_{l,w} n(w, l)P(z, r|w, l)}{\sum_{w,l} n(w, l) \sum_z P(z, r|w, l)} \quad (10)$$

$$P(r) = \frac{\sum_{w,l} n(w, l) \sum_z P(z, r|w, l)}{\sum_{w,l} n(w, l)} \quad (11)$$

$$\mu_r = \frac{\sum_{w,l} n(w, l) \sum_z P(z, r|w, l)l}{\sum_{w,l} n(w, l) \sum_z P(z, r|w, l)} \quad (12)$$

$$\Sigma_r = \frac{\sum_{w,l} n(w, l) \sum_z P(z, r|w, l)(l-\mu_r)(l-\mu_r)^{\mathsf{T}}}{\sum_{w,l} n(w, l) \sum_z P(z, r|w, l)} \quad (13)$$

The E-step and the M-step equations are alternated until a convergence criteria is met.

[3] In the following, we omit the pedices for clarity in the reading, resorting them only when necessary.

After the learning, several useful inferences can be assessed. The geo-topic membership label of an image can be estimated in a Maximum Likelihood framework: starting from the image likelihood $\sum_w n(w, l_i) \log p(w, l_i)$, we build the geo-topic class conditional, calculating then the geo-topic index $R(l_i) \in 1,...,R$ that maximizes it. In formulae

$$\begin{aligned} R(l_i) &= \arg\max_c \sum_w n(w, l_i) \log p(w, l_i|r_c) \\ &= \arg\max_c \sum_w n(w, l_i) \log \sum_z P(w|z)P(z|r_c)p(l_i|r_c) \end{aligned} \tag{14}$$

The inference for the geo-recognition consists in setting the location of an image as being the center of a region, for all the regions discovered. Then, the learning algorithm is run, leaving locked the $p(w|z)$ distribution, *i.e.*, the statistical word-descriptions of each visual topic, and the parameters μ_{r_c}, Σ_{r_c}, estimating thus $p(z|r_c)$. At the end of the process, we obtain a set of distributions $\{\hat{p}(z|r_c)\}$ for $r_1,..., r_R$. Each one of these distributions is compared with the correspondent training distributions $\{p(z|r_c)\}$, employing the Resistor-Average distance [3], *i.e.*, a symmetric similarity score, based on the Kullback-Leibler divergence $KL(\cdot||\cdot)$; in formulae:

$$\mathcal{R}(p, q) = [KL(p||q)^{-1} + KL(q||p)^{-1}]^{-1} \tag{15}$$

The region of the un-labelled image $\hat{d}$, *i.e.*, $\hat{r}$ is the one that satisfies the following equation:

$$\hat{r} = \arg\max_c \mathcal{R}(p(z|r_c), \hat{p}(z|r_c)) \tag{16}$$

It is worth noting that the region recognition works also in the case of multiple un-located images, with the hypothesis that all the images come from the same region.

3.2. Comments about LD-pLSA

This section aims at highlighting the differences of our technique with respect to existent pLSA extensions.

- Spatial layout analysis techniques [6, 11, 12, 13] - these techniques learn 1) the locations associated to *visual topics* in the images [11]; 2) the locations of the *words of a single visual topic* grouping them in a single cluster [6, 12], also introducing robust management of the clutter [13]. Conversely, our purpose is to model the *locations of the images* in a geographic area.

- Topological pLSA [8] - this technique adds an additional latent variable, which lives in the same space of the original topic variable z, and serves to capture similarity relations among topics. Even if the form of the joint probability distribution is similar to ours (see pag.167 in [8]), its parametrization and its meaning is totally different.

4. Experimental results

4.1. Dataset details

To analyze our framework, we built a geo-located image database crawling 3013 images from Panoramio, considering the southeastern part of France. The download has been accomplished considering all the geographical zones where images were present, in an uniform way. We chosen France because of its large variety of natural scenes, ranging from mountains to sea areas, with historical, industrial or coastal cities, fields and villages. Please note that, at the best of our knowledge, no public dataset of geo-located images is available for experimental evaluations; image URL addresses used for the experiments are listed here[4]. As preprocessing, we converted all the images to grayscale, and we resized them via bicubic interpolation to standard width (320 pixels). Then, we employed the difference of Gaussians (DoG) point detector [14], which selects sparse blob-like patches invariant to translation, rotation, scale, and uniform illumination variations. We chose this solution, disregarding to adopt affine-invariant detectors [16, 21, 15, 9], well-suited for a object recognition context, as discussed in [17]. Once interest points are extracted (averagely 670 for each image), 128-dim. SIFT descriptors [14] are applied to describe their local neighborhood. Subsequently, descriptors are quantized into a codebook of $M = 300$ visual words via k-means, where low populated clusters (≤ 10 elements) are pruned out. Once the visual words are estimated, *bags of words* (BOV) representations are built for each image.

The LD-pLSA training needs to set two parameters: the number of visual topics Z and the number of geo-topics R. For what concerns Z, we start choosing $Z = 32$, considering the analysis made in [17]. Then, we set the number of geo-topics $R = 16$. The choice of this parameter depends on the level of detail that an user wants to achieve in analyzing a particular geographical area. We will go back later on this issue. A principled (not only exploratory) model selection issue on the "best" Z and R has not been considered in this paper, but it will be subject of future research.

We run the EM algorithm, that takes approximately 3 minutes to converge after 300 iterations on a 1.98 Ghz Xeon with a Matlab implementation. After the learning, we obtain a set of Gaussian geo-topics, simply called *regions*, each one modeled by a distribution over visual topics $P(z|r_c)$, where the visual topics are co-occurrences of visual words described by $P(w|z_k)$.

4.2. Geo-categorization

The Gaussian regions are depicted in Fig.2a, and the image locations, labeled as described in Eq.14, are shown in Fig.2b. For each region, in Fig.2 on the right, we

[4] http://profs.sci.univr.it/∼cristanm /research/cvpr08.html

show some of the member pictures, drawn from left to right in decreasing order of geo-topic conditional likelihood $\sum_{w} n(w, l_i) \log p(w, l_i|r_c)$ (see Eq.14); the last two column shows images with very low conditional likelihood. It is easy to see how each of the discovered regions contains images which are visually related. In particular, it is worth to notice how each region individuates a particular *scene category*, in the sense described in [22]: actually, we see cities (reg. 1,3,5,14,17,18,19), (mostly) fields (reg. 2,4), mountains (reg. 6,8,12,13), mountain villages (reg.15), coastal areas (reg.7,9,16), lakes (reg.20). This can also be noticed by looking at Fig.2a; actually each category lies entirely on a particular kind of landscape (mountains, plains, coasts). In this sense, we can say that the categories built are *geographically coherent*.

4.2.1 Details on the geo-topics

In order to probe how the geo-topics are characterized, for each visual topic z_k we rank the M visual words in decreasing order of $P(w_j|z_k), j = 1,..., M$. Then, we select the first six visual words (that we address as *representative*) of all the visual topics. For some member images of all the geo-topics, we show the location of the representative words of the visual topic that best models that region, *i.e.*, for a given region r_c, we extract the visual *typical* topic index k for which $P(z_k|r_c)$ is maximum (see Fig.3). Actually, for almost each region there is a visual topic which strongly characterizes its most prominent visual aspect. For example, in the region 1 (Avignon), representative visual words are located on the curved architectural elements of the bridge and the palace, and not on the vegetation.[5]. Viceversa, on region 2 (Val de Durans) and region 4 (Provence plains), visual words are mostly located on the vegetation, disregarding the building, which are present in a lower quantity on the images of those classes.

In region 9 (Arles, Camargue National Park), being present flat swamp and wild beach coasts, visual words are mostly located on the vegetation and on the horizon line. On region 13 (Alpes da Haute Provence) many pictures depict curved hill profiles and vegetation, as highlighted by the representative words shown. Further, regions depicting mountains exhibit recurrent rock patterns distilled by the visual words. It is worth to notice how visual topic 11 is typical for the cities.

As additional exploration of the *visual* content encoded in the conditional $P(z_k|r_c)$, given a test image, we visualize the representative visual words of a topic, which is the typical for a region different to that of the test image. As visible in Fig.4a, the test image comes from the region 17 (Montecarlo). At first, we select the topic $z = 20$, typical for region 4 (Provence fields, see one of these images on top of Fig.4a), whose representative words are mostly located on vegetation patterns. As one can observe, in the test image very few visual words are present. We then select topic $z = 13$, typical for region 1 (Avignon), in which many words lie on the (curved) architectural elements of the ancient buildings present there. In the test image, interestingly, visual words result on locations mimicking the Avignon's architectural patterns. Similar considerations hold for several images in the dataset.

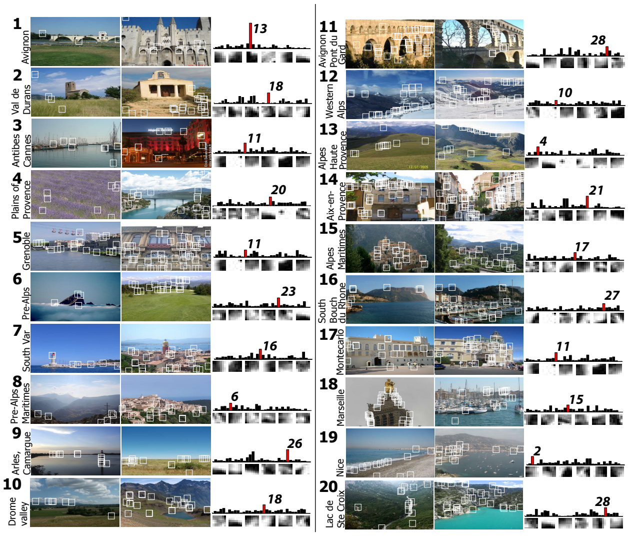

Figure 3. Topic representations: for each region we highlight the most prominent *typical* visual topic (numbered in red on the respective $p(z|r)$ histograms). The positions of the 6 most representative visual words of the typical topic are shown as rectangles on the images on the right; below the histograms, such words are visualized as squared patches.

Another means to highlight the peculiarity of the geo-topic descriptions consists in building a pairwise dissimilarity matrix among the $p(z|r)$ distributions. As dissimilarity measure between distributions, we employ again the Resistor-Average distance. The dissimilarity matrix, depicted in Fig.4b, shows in general an high degree of nonuniformity among the $p(z|r)$, which will be useful in the recognition step. More in detail, performing hierarchial clustering on the dissimilarity matrix and visualizing the resulting dendrogram (Fig.4c), we discover the tendency of cities and landscape regions to forming separated macro-groups with consistent intra-group differentiation. Regions 1,7,and 9 (Avignon, South Var, and Camargue, respectively) are strongly differentiated by the rest of the regions.

Finally, we test our categorization technique by varying the number C of geo-topics (see Fig. 5a,b,c). As written before, a large number of geo-topics permits to describe more finely the different geographical areas spanned by the images. Anyway, augmenting the number of geo-

[5] To better highlight this fact, we calculate a patch representation for each visual word, following the approach described in [10], showing it on the right of the images.

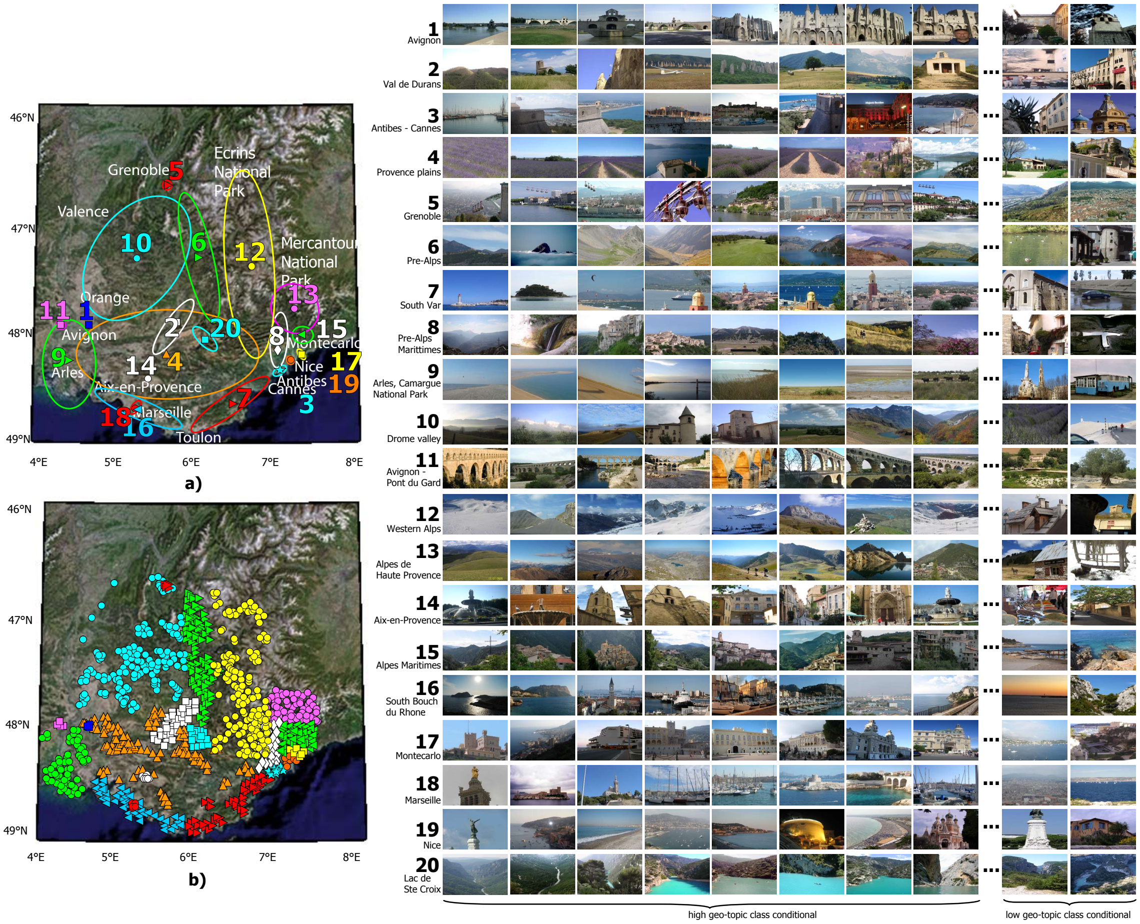

Figure 2. Categorization results: a) (numbered) Gaussian regions discovered by LD-pLSA. b) labelled image locations. On the right, for each region a set of 10 member images: 8 have high region-membership conditional likelihood, while the last two have lower conditional likelihood.

topics causes the creation of geo-categories formed by few member images. As we see in the following, this brings to bad geo-recognition performances. Here we visualize how our method behaves fixing the number of geo-topics to $C = 2, 3, 4$ (Fig.5 a,b,c respectively). Even with $C = 3$, the regions portrayed define consistent geographical areas. Finally, changing of the number of visual topics K, from 16 to 100, does not affect the quality of the obtained partitions.

4.2.2 Comparative results on geo-categorization

In order to assess the relevance of our geo-categorization technique, we set up two alternative strategies to partition our dataset: 1) **Location dependent strategy** - performing Gaussian clustering via EM, considering only the location information of the images; 2) **Aspect dependent strategy** - performing classic pLSA with $T = 32$ topics on the dataset, obtaining as image feature vectors the distributions $p(z|d)$ (different for each image d); subsequently, performing Gaussian clustering via EM using such features vectors.

In the location dependent strategy, as expected, several incoherent geo-categories (*i.e.*,whose member images exhibit strongly different visual patterns) are produced (see Fig. 5d); in the aspect dependent strategy, we have very sparse and overlapped small clusters, not depicted here for clarity. Anyway, note how locations of pictures portraying cities exhibit the same geo-category label (Fig. 5e).
As comparative categorization strategy employing both location and visual aspect information, we perform non-parametric clustering via Mean Shift [4], employing a product kernel ([4], pag.610) on the feature vectors obtained by concatenating the locations and the latent distributions gathered through pLSA.

This strategy gives better results than the first two alternative policies, but behaves worse with respect to our method (see Fig. 5f: the method is not able to discover the Cote d-Azur cities). Actually, LD-pLSA partitions the data

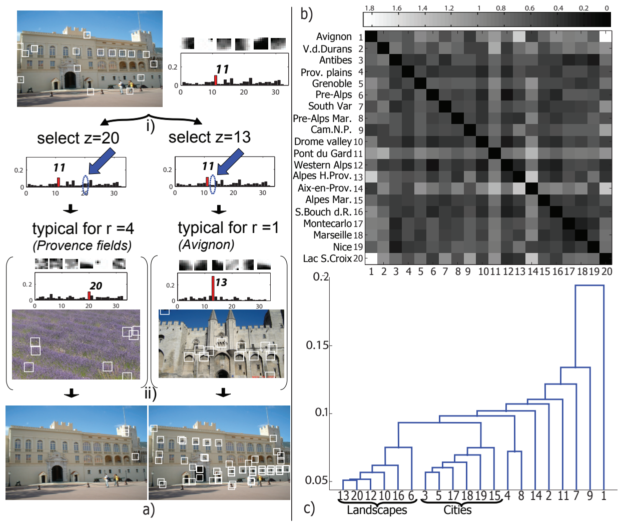

Figure 4. Experiments on geo-topics: a) visualization of representative words of different topics on a test image; b)dissimilarity matrix among $p(z|r)$ distribution; c) dendrogram resulting from the clustering of the dissimilarity matrix.

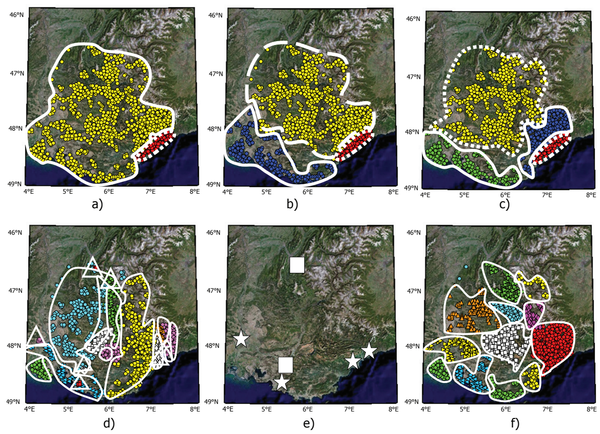

Figure 5. Geo-categorization results: LD-pLSA with a) C=1, b) C=2, c) C=3. Comparative strategies: d) Location dependent strategy; e) Aspect dependent strategy; f) Mean Shift based strategy on combined data location+visual topic information.

into groups and select the features to index these groups (i.e., $p(z|r)$ geo-topic distributions) at the same time, during the EM training. Instead, the Mean-Shift strategy works *a posteriori*, separating the images into groups, resorting to the extracted individual image features (i.e., $p(z|d)$ distributions) ignoring the location aspect.

4.3. Geo-recognition

The geo-recognition task consists in inferring the region, found by previous geo-categorization, depicted by image(s) with no available location information. In order to evaluate our strategy, we split our dataset into training and testing sets; the training set is used to geo-categorize the images. We employ now the categorization shown in Sec.4.2, with $C = 20$ regions. Together with the testing set, we build also the correspondent ground truth label set, associating to each test image the label of the geo-category in which it is located. We vary the size and the content of training and testing sets, building a Leave-K-Out (LKO) strategy [18] to cross-validate the results. We select $K = 50$, obtaining 50 subsets; 49 subsets form the training set, the 50-th is the testing set. We pay attention that each subset is formed by images covering uniformly the entire geographic area spanned by the whole geo-located image dataset.

It is worth to note that the geo-recognition mechanism works also with more than one image, with the only constraint that all the images comes from a single place. Therefore, for each testing set, we variate the number of testing images used as a single query, in a way such that each region could be evaluated by an increasing number of pictures. Then, the final results of recognition are mediated on all the 50 runs of the LKO strategy. We call this strategy *Simultaneous LD-pLSA* (**S-LD-pLSA**).

As comparative strategy for the recognition, we used the Support Vector Machine (SVM)-based strategy proposed in [17], in which multi-classes SVMs were trained on the BOV representation of the member images of each category. Unfortunately, here it is not possible to evaluate directly the effect of augmenting the number of images for a single query. Therefore, in order to make an appropriate comparative evaluation, we decide to adopt a majority criteria for both our framework and the SVM approach: in practice, given a query with I images, we evaluate the response of our classifier for all the I images taken separately, considering as winner the region which gains the biggest number of individual votes. We call these serial strategies for our approach and the SVM approach as *Majority LD-pLSA* (**M-LD-pLSA**) and *Majority SVM* (**M-SVM**), respectively. The results are shown in Fig.6a, obtained considering queries formed by 1, 5 , 10, 20 and 50 images respectively. As we can see, M-SVM outperforms M-LD-

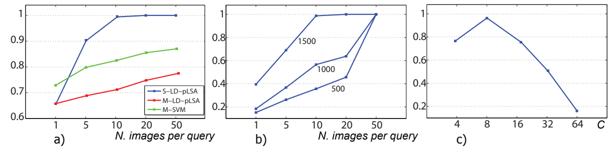

Figure 6. Recognition tests: a) classification accuracy; b) classification accuracy with training sets of different cardinalities (indicated under the curves); c) classification accuracy in the case of queries formed by a single image, varying the number C of geo-topics considered in the categorization step.

pLSA and S-LD-pLSA in the case of single-image query. Anyway, S-LD-pLSA outperforms M-SVM when queries are formed by multiple images. In Fig.6b, different classification accuracy curves are reported, obtained by employing training sets for geo- categorization of different cardinalities, extracted by the whole dataset and cross validated as explained before. As expected, diminishing the training set, the performances degrade. Note that the different training sets cover uniformly the entire geographic area spanned by the original dataset. Finally, we evaluate the classification accuracy when varying the number C of regions in the geo-categorization step, again adopting the LKO strategy explained above on the entire dataset. Here we do not have a regular monotonic decreasing slope (Fig.6c): actually, a low number of geo-topics in the geo-categorization gives rise to regions not well characterized. Vice versa, a high C means several regions with few member images.

5. Conclusions

In this paper, we focus on geo-located images, *i.e.*, images whose acquisition location is given. In such framework, we focus on two novel issues, which are 1) how to categorize geo-located images considering both spatial and visual information, and 2) how to infer the zone surrounding the location of acquisition of an image non geo-located. To this end, we developed a novel statistical technique, based on probabilistic Latent Semantic Analysis. The proposed technique characterizes the entire area spanned by the geo-located images as a set of latent geographical regions characterized by a distribution of latent visual aspects. The visual aspects of each region are thoroughly characterized by distilling recurrent visual patterns captured via bag of words representations. At the same time, the technique is able to infer the region of one or more images portraying the same unknown location. Comprehensive experiments show the pro and cons of the proposed technique. Future developments of the model will regard model selection issues, and how to describe a location under different levels of detail.

References

[1] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, 2003.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation, 2003.

[3] N. Checka and K. Wilson. Symmetrizing the kullback-leibler distance. Technical report, Rice University, 2002.

[4] D. Comaniciu and P. Meer. Mean Shift: A Robust Approach Toward Feature Space Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, 2002.

[5] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39:1–38, 1977.

[6] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google"s image search. In *ICCV '05*, pages 1816–1823, Washington, DC, USA, 2005. IEEE Computer Society.

[7] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99*, pages 50–57, New York, NY, USA, 1999. ACM Press.

[8] T. Hofmann. Probabilistic topic maps: Navigating through large text collections. In *IDA '99*, pages 161–172, London, UK, 1999. Springer-Verlag.

[9] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *European Conference on Computer Vision*, pages 228–241, 2004.

[10] F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR '05*, volume 2, pages 524–531, Washington, DC, USA, 2005. IEEE Computer Society.

[11] D. Liu, D. Chen, and T. Chen. Latent layout analysis for discovering objects in images. In *ICPR '06*, pages 468–471, Washington, DC, USA, 2006. IEEE Computer Society.

[12] D. Liu and T. Chen. Semantic-shift for unsupervised object detection. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 16, Washington, DC, USA, 2006. IEEE Computer Society.

[13] D. Liu and T. Chen. Unsupervised image categorization and object localization using topic models and correspondences between images. In *ICCV '07*, Washington, DC, USA, 2007. IEEE Computer Society.

[14] D. Lowe. Object recognition from local scale-invariant features. In *ICCV '99*, volume 2, pages 1150–1157, 1999.

[15] J. Matas, O. Chum, U. Martin, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British Machine Vision Conference*, volume 1, pages 384–393, 2002.

[16] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[17] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars. A thousand words in a scene. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1575–1589, 2007.

[18] J. Shao. Linear model selection by cross-validation. 88(422):486–494, 1993.

[19] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[20] K. Toyama, R. Logan, and A. Roseway. Geographic location tags on digital images. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 156–166, New York, NY, USA, 2003. ACM.

[21] T. Tuytelaars and L. V. Gool. Matching widely separated views based on affine invariant regions. *Int. J. Comput. Vision*, 59(1):61–85, 2004.

[22] J. Vogel and B. Schiele. A semantic typicality measure for natural scene categorization. In *DAGM-Symposium*, pages 195–203, 2004.