Geo-located image categorization and location recognition

Marco Cristani[*†] Alessandro Perina[‡*] Umberto Castellani[§*]

Vittorio Murino[¶*]

Abstract

Image categorization is undoubtedly one of the most recent and challenging problems faced in Computer Vision. The scientific literature is plenty of methods more or less efficient and dedicated to a specific class of images; further, commercial systems are also going to be advertised in the market. Nowadays, additional data can also be attached to the images, enriching its semantic interpretation beyond the pure appearance. This is the case of geo-location data that contain information about the geographical place where an image has been acquired. This data allow, if not require, a different management of the images, for instance, to the purpose of easy retrieval from a repository, or of identifying the geographical place of an unknown picture, given a geo-referenced image repository. This paper

[*]University of Verona, Computer Science Department, Strada Le Grazie 15, 37134 Verona, Italy

[†]marco.cristani@univr.it, Tel. +39 045 8027988, Fax +39 045 8027068

[‡]alessandro.perina@univr.it, Tel. +39 045 8027803

[§]umberto.castellani@univr.it, Tel. +39 045 8027988

[¶]vittorio.murino@univr.it, Tel. +39 045 8027996

constitutes a first step in this sense, presenting a method for geo-referenced image categorization, and for the recognition of the geographical location of an image without such information available. The solutions presented are based on robust pattern recognition techniques, such as the probabilistic Latent Semantic Analysis, the Mean Shift clustering and the Support Vector Machines. Experiments have been carried out on a couple of geographical image databases: 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 [1]. Unfortunately, such problem is very hard at least for two reasons: first, because the meaning of a picture is an ephemeral entity, extrapolated subjectively by human beings; the second reason is the semantic gap, *i.e.*, the gap between the object in the world and the information in a (computational) description derived from a recording of that scene [1]. Despite this, the image categorization research field is one of the most fertile area in Computer Vision: an interesting, even if dated, review can be found in [1] , where a taxonomy of the main algorithms for image categorization and retrieval is presented. In [2], a comprehensive survey of the public available retrieval systems is reported, and challenges and some future perspectives for the retrieval systems are discussed in [3].

The common working hypothesis of most categorization algorithms is that images are located in a single repository, and described with features vectors summarizing

their visual properties. Recently, this classical framework has been improved with the use of textual labels or *tags*, associated to the images. Textual labels are usually given by a human user in order to constrain the number of ways an automatic system can categorize an image, and suggest to the viewers the information the author of the picture wants to communicate with it.

Very recently, this framework has been further updated with the introduction on the market of several cheap GPS devices, mounted on the cameras. Such devices automatically assign tags to the captured pictures, indicating the geographical position of the shot. This capability charmed researchers and web designers, which understood the potential scenario of a novel and more advanced way of sharing pictures, succeeding and outperforming the "non-spatial" public image databases. This caused the creation of global repositories for the *geo-located* images, as in Panoramio[1], and the addition of novel functionalities for the display of geo-located images in Google Earth[2] and Flickr[3]. More specifically, the interfaces for the visualization of geo-located pictures of Google Earth and Flickr insert over the satellite maps particular icons that indicate the presence of a picture taken in that place, that the user can click over and enlarge. The interface of Panoramio, exclusively suited for the maintenance of geo-located pictures, is more structured. Pictures are visualized as thumbnails on a side frame, representing the images geo-located on a satellite map. These interfaces allow to effectively exploit geographical tags, permitting the users a novel way to discover places, more personal and emotional.

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

[2] `http://earth.google.com/`

[3] `http://www.flickr.com/`

As we will see in the following, this new framework discloses an innumerable set of novel and stirring applications, that go beyond the mere visualization, which have to be carefully explored by the researchers, and poses novel problems to be faced in the realm of the image categorization. In this paper we analyze two of these applications.

The first *underlies* and ameliorates the management and visualization of the geo-located images. In all the interfaces, the exploration of a geo-located image database occurs by zooming on a map in a desired location and visualizing a set of randomly sampled images lying in the related neighborhood. This layout becomes very unattractive and uncomfortable in presence of a massive number of images, as it is currently in all the databases considered. As a solution, an effective way to *categorize* geo-located images has to be proposed, in which images have to be clustered together by taking into account, other than the associated visual properties, also the geographical position of acquisition.

In this way, the exploration of a geo-located database can be strongly improved. Grouping the images for similarity and proximity permits to create *geo-clusters* from which a small number of representative images can be extracted and visualized. In this way, a better global visualization scheme can be exploited, in which each depicted picture represents a different geographical pattern; in other words, each different zone depicted on the map can be visualized by means of few good representatives.

Another interesting and harder issue to be dealt with is the *geo-location* of images, where the goal is to infer the geographical zone in which a picture not geo-tagged has been acquired. This is useful in a entertainment context, in which one want to fill his geo-located image database with non-tagged photos. Another context could be the

forensic one, where it results essential to constrain the possible zone in which a picture has been taken.

A similar issue was faced few years ago, under the name of *location recognition task*, as an open research contest [4]. There, contestants were given a collection of color images taken by a calibrated digital camera. The photographs had been taken at various locations taken in a *small* city neighborhood, often sharing overlapping fields of view or certain objects in common. The GPS locations for a subset of the images are provided. The goal of the contest was to guess, as accurately as possible, the GPS locations of the unlabeled images. Essentially, all the proposed resolutive approaches were based on the reconstruction of 3D scenes owing to the registration of several images with overlapping fields of view. Inferences on the position of non geo-located test images was inferred by taking into account that 3D model. An example of such framework is proposed in [4].

In our situation, the task is much harder: heterogeneous pictures taken far from each other, at a different time of the day, have to be managed. This is a difficult problem and, to the best of our knowledge, no solutions are present nowadays. Due to the vastity of the existent geographical varieties, it seems now 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 face the issues of the *geo-clustering* and geo-location recognition of images, in the context of a large geo-located image database. We will show how

[4] *Where Am I?* ICCV Computer Vision Contest, please see

`http://research.microsoft.com/iccv2005/Contest/`

using well-known techniques in the literature, such as the Probabilistic Latent Semantic Analysis, Mean Shift Clustering and Support Vector Machine framework, strong and effective results can be achieved, proposing valuable solutions to the problems discussed above.

The rest of the paper is organized as follows. In Sec. 2, mathematical background notions are reported. Then, in Sec. 3, the outline of our system for geo-clustering and geo-location recognition is detailed. Sec. 4 presents the experiments carried out on large databases taken from Panoramio, and, finally, Sec. 5 concludes the paper, envisaging future perspectives.

2 Mathematical background

2.1 Probabilistic Latent Semantic Analysis

In this section, we briefly review the probabilistic Latent Semantic Analysis (pLSA), in its adaption to visual data. We describe the model using the classical terminology of the literature on text classification, in parallel to that regarding the image domain. The input is a dataset of D documents (images), each containing local regions found by interest operators, whose appearance has been quantized into W visual words [5]. Therefore, the dataset is encoded by a co-occurrence matrix of size $W \times D$, where the location $< w, d >$ indicates the number of (visual) words w in the document d. The model incorporates a single latent topic variable, z, that links the occurrence of word

w to document d. In formulae:

$$P(w,d) = \sum_{z=1}^{Z} P(w|z)P(z|d)P(d) \tag{1}$$

As a result, we have obtained a decomposition of a $W \times D$ matrix into a $W \times Z$ matrix and a $Z \times D$ one. Each image is modeled as a probability distribution over the topics, *i.e.*, $P(z|d)$; the distribution $P(w|z)$ encodes the topic z, as a probabilistic co-occurrence of words. The distributions of the model, $P(w|z)$ and $P(z|d)$, are learnt using Expectation Maximization (EM) [6]. The E-step computes the posterior over the topic, $P(z|w,d)$ and then the M-step updates the densities. This maximizes the likelihood L of the model over the data:

$$L = \prod_{d=1}^{D} \prod_{w=1}^{W} P(w,d)^{n(w,d)} \tag{2}$$

In recognition, the distribution $P(w|z)$ is locked and EM is applied, estimating the $P(z|d)$ for the query images. For a deeper review of pLSA, see [7]; for an application on scene recognition, see [8].

2.2 Mean Shift clustering

The Mean Shift (MS) procedure is an old, recently re-discovered non-parametric density estimation technique [9, 10]; the theoretical framework of the MS arises from the Parzen Windows technique [11], that under particular hypotheses of regularity of the input space (independency among dimensions, see [10] for further details) estimates the density at the $d-$dimensional point $\mathbf{x}$ as:

$$\hat{f}_{h,k}(\mathbf{x}) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^{n} k\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right) \tag{3}$$

where $c_{k,d}$ is a normalizing constant, n is the number of points available, and $k(\cdot)$ the kernel *profile*, that models how strongly the points are taken into account for the estimation in dependence with their distance h (also called kernel *bandwidth*) to $\mathbf{x}$.

Mean Shift extends this "static" expression, differentiating (3) with respect to $\mathbf{x}$ and obtaining the density gradient estimator

$$\nabla \hat{f}_{h,k}(\mathbf{x}) = \frac{2c_{k,d}}{nh^d}\left[\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}_i - \mathbf{x}}{h}\right\|^2\right)\right]\left[\frac{\sum_{i=1}^{n}\mathbf{x}_i g\left(\left\|\frac{\mathbf{x}_i-\mathbf{x}}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}_i-\mathbf{x}}{h}\right\|^2\right)} - \mathbf{x}\right] \tag{4}$$

where $g(x) = k'(x)$. This quantity is composed by three terms: the first is a normalizing constant, the second one in square brackets is *proportional* to the normalized density gradient obtained with the kernel profile k and the third one is the *Mean Shift* vector, that is guaranteed to point towards the direction of maximum increase of the density. Therefore, starting from a point $\mathbf{x}_i$ in the feature space, and applying iteratively the MS vector, a trajectory is produced which converges on a stationary point $\mathbf{y}_i$, representing a mode of the whole feature space.

The MS procedure is widely applied to clustering issues: the first step of the clustering is made by applying the MS procedure to all the points to be analyzed $\{\mathbf{x}_i\}$, producing several convergency points $\{\mathbf{y}_i\}$. A consistent number of close locations of convergence, $\{\mathbf{y}_i\}_l$, indicates a mode μ_l. The labeling consists in marking the corresponding points $\{\mathbf{x}_i\}_l$ that produced the set $\{\mathbf{y}_i\}_l$ with the label l. This happens for all the convergence locations $l = 1, 2, \ldots, L$.

3 The proposed method

Given our set of geo-located images, the first step toward the geo-clustering consists in deriving a high level representation of the visual content of such images, without relying on the geo-locations. This is achieved by the topic representation of the images given by pLSA. Specifically, affine elliptical regions are estimated for each image converted in grey scale, constructed by elliptical shape adaptation about an interest point [12]. Each region is mapped to a circle by appropriate scaling along its principal axis and a 128-dim SIFT descriptor is built. Then, descriptors are quantized into visual words via K-means, and histogram word representations are built for each image. Finally, the topic representation is obtained via EM.

Now, each image is described by a point in a $Z-$dimensional *topic* space. Adopting an Euclidean distance and performing clustering in this space would group visually similar images. At this point, we augment the image description by adding, for each image, the related geo-locations. In this way, we move in an *augmented* space, formed by the topic subspace and the *geographical* subspace, that we suppose for convenience as independent. In other words, each image d is described with a feature vector $[P(z|d), g(d)]$, where $g(d) \in \mathcal{R}^2$ is a couple containing its latitude and longitude values.

In order to perform clustering in the augmented domain, a multivariate kernel profile is used [10], that is:

$$k(\mathbf{x}) = \frac{C}{h_z^Z h_g^2} \prod_{u \in \{z,g\}} k\left(\left\| \frac{\mathbf{x}_u}{h_u} \right\|^2 \right) \tag{5}$$

where C is a normalization constant, and h_z,h_g are the kernel bandwidths for the topic

and the geographical sub-domain, respectively. This kernel is the product of two intra-subspace kernels, and it weights in a different way each subspace, depending on the kernel bandwidth associated.

As intra-subspace kernel $k(\cdot)$, we adopt the Epanechnikov kernel [10], that differentiated (see Eq.4) leads to the uniform kernel $g(\cdot)$, *i.e.*, a multidimensional unit sphere.

The choice of the number of topics and the values for the bandwidths is an aspect discussed in the next section. After the clustering, we obtain a set of classes which represent particular compact zones containing images with similar appearance.

The second task, i.e. the geo-location recognition, is achieved by employing the Support Vector Machines (SVMs) [13]. SVM constructs a maximal margin hyperplane in a high dimensional feature space, by mapping the original features through a kernel function. Here, a SVM classifier with Radial Basis Functions (RBF) has been trained to discriminate the clusters obtained via pLSA and MS clustering. In the SVM training, the geographical features of the images of the different clusters are discarded, being our task the geo-location recognition, *i.e.*, after the training we need to operate on features vectors in which the geographical information is not provided.

Then, for a novel image of unknown geo-location, we estimate its topic distribution locking the $P(w|z)$ estimated on all the data via pLSA and running the EM algorithm (see Sec. 2.1). The obtained distribution is fed as input in the SVM classifier, which has been employed in a multi-class framework, by adopting the *one-against-one* policy[14]. As a result, we obtain the label of the region which the input image likely belongs to.

4 Experiments

To validate our framework, we built two databases considering the Hawaii Big Island (*Hawaii* database), and the southern part of France (*France* database, see Fig.1).

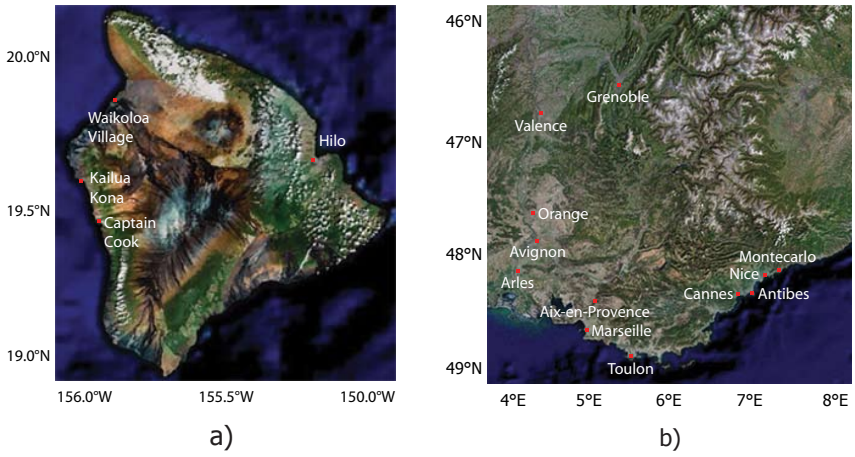

Figure 1: Geographical zones analyzed: a) Hawaii Big Island, b) Southern France

The databases are composed by 1013 and 607 geo-located pictures, respectively, downloaded from Panoramio. We choose Hawaii Big Island because of its large variety of natural scenes, ranging from mountains to sea, with volcanos, cascades and villages. Similar considerations hold for the France database.

At first, we perform pLSA analysis, using $Z = 15$ topics in both the databases. Then, we perform Mean Shift clustering adopting a multivariate kernel (with bandwidth values equal to $h_z = 0.3$ for the topic space and $h_g = 0.2$ for the geographic space[5]. The obtained results can be observable in Fig.2 and Fig. 3.

[5]Regarding the parameters, changing the number of topics (we try $Z = 4, \ldots, 30$) does not modify drastically the quantity and the nature of the clusters obtained. The choice of the kernel bandwidths is not critical, and easy to set.

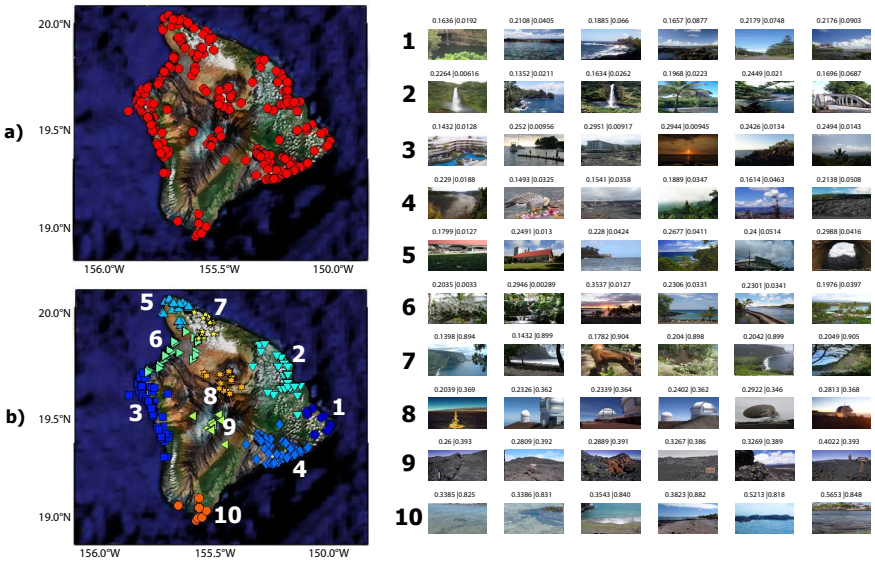

Figure 2: Hawaii database: a) location of the input photos; b) clustering results; on the right, member images of each cluster depicted in b) are shown. On top of each image there is the Euclidean distance between its topic distribution and the topic distribution of the cluster centroid (left); on the top-right, the Euclidean distance between its location and the location of the centroid.

Together with the input datasets (part (a) of each figure), the clustering results (part (b) of each figure), in the figures we show for all the clusters discovered, some member photos depicted in ascending order w.r.t a similarity measure relative to the centroid of the cluster. Such measure is the Euclidean distance between the topic representation of an image and that of the centroid, multiplied by the geographical Euclidean distance. The value of both the sub-distances are attached over the photos.

As visible in Fig.2, our clustering procedure is able to separate geographically close

zones, such as the zones 5, 6, 7, which exhibit different recurrent visual patterns (zone 5 - flat coasts with buildings; zone 6 - wild beaches; zone 7 - high rocky coasts). The zone 3 is mostly formed by vegetation and cascades, zone 8 and 9 lie upon the volcanos and zone 1, 4, and 10 represent flat coasts, volcanic areas facing the sea and rocky coasts, respectively.

Similar considerations hold for the France database. In Fig.3a the location of all the images are shown. In Fig.3b the clustering results are shown. In the clustering, we apply a size filter to discard clusters with less than 5 images. For this reason, some of the original image locations are not shown in Fig.3 b.

In this database, the capabilities of our clustering framework are even more highlighted: compact groups of images on the map are separated, representing highly different visual patterns. For example, in zone 3, we can see Montecarlo; zone 5 comprehends Cannes-Antibes. Other clusters are: zone 9 - Avignon, zone 10 - Arles, zone 11 - Pont du Gard, zone 12 - Parc Naturel de Camargue.

In order to investigate on the value added by coupling visual similarity and proximity relation, we perform Mean Shift clustering of the images of the France dataset a) by taking into account only the geographical position, and b) only considering the topic distribution (Fig.4a and b, respectively), employing the same correspondent bandwidth values adopted in the proposed method.

In the clustering performed by considering only the spatial subdomain, groups of photos related to visually different geographical zones are fused together, as occurred for clusters 10 and 12, and clusters 5 and 3 (see Fig. 3). In the clustering based only on topic information, the clusters are sparse and spread out over the entire map.

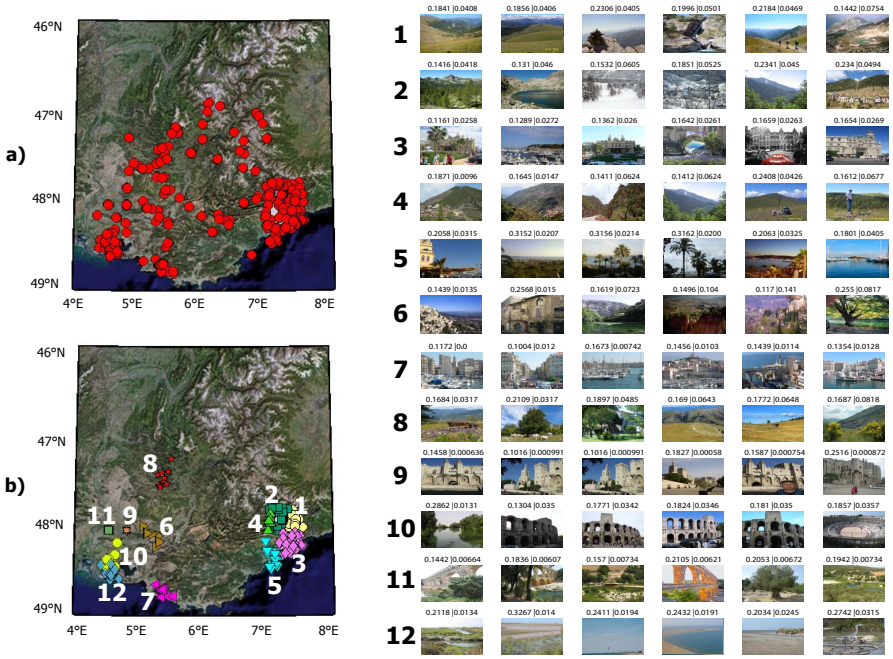

Figure 3: France database: a) location of the input photos; b) clustering results; on the right, member images of each cluster depicted in b) are shown. On top of each image there is the Euclidean distance between its topic distribution and the topic distribution of the cluster centroid (left); on the top-right, the Euclidean distance between its location and the location of the centroid.

Here, it is worth to note that the cluster depicted by yellow stars represent two cities, Cannes and Marseille, which are geographically far but visually comparable. Similar considerations hold also for the other clusters.

We perform the same test with the Hawaii database, obtaining similar results not shown here due to the lack of space.

For what concerns the recognition task, since the Radial Basis Function (RBF)

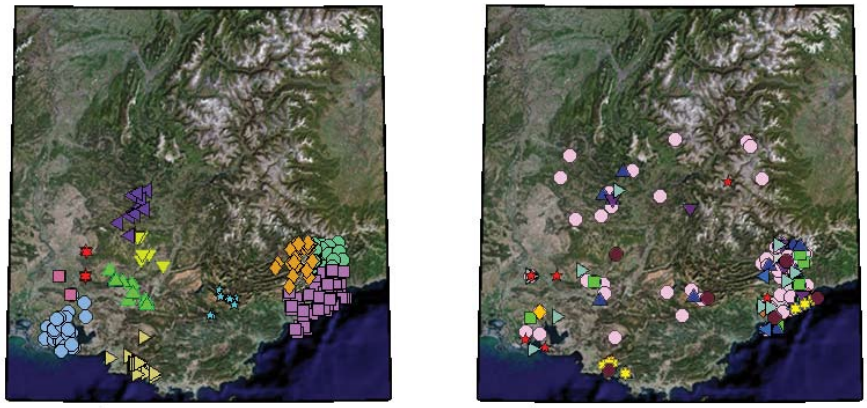

Figure 4: Clustering results by considering: a) only geographical information; b) only topic information

kernel has been used, two parameters C and γ needed to be estimated. According to suggestions reported in [15], data are normalized properly and parameters are estimated by combining grid search with leave-one-out cross-validation [11]. In order to extend the SVM to a multi-class framework, the one-against-one approach is carried out [11]. We obtain 85.24% of the accuracy on the Hawaii database, and 75% on the France database. In this way, an unknown picture can be located in the right geo-location, with an uncertainty given by the area of the selected cluster: the larger the cluster, the more uncertain is the exact location where a picture has been taken.

5 Conclusions

In this paper we propose a framework that faces successfully two novel and promising applications in the image categorization realm, which are the geo-clustering and

the geo-location recognition. Geo-clustering consists in group together images which are 1) visually similar and 2) taken in the same geographical area. This application serves for a more effective management and visualization of geo-located images, *i.e.*, images provided with geographical tags, indicating the location of the acquisition. Geo-location recognition consists in inferring the geo-location of a picture whose provenance is unknown, with the help of a geo-located image database. The solutions proposed with our framework employ robust pattern recognition techniques, such as probabilistic Latent Semantic Analysis, Mean Shift clustering and Support Vector Machines. This work indicates a set of future perspectives to be investigated. For example, we are currently studying a way to create of an high level description for geo-located images, such as the one provided by the pLSA, which incorporates also the location in which the picture has been taken. Moreover, we are studying a multi-level description, able to increase the geographical precision with which an image can be geo-located.

References

[1] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, "Content-Based Image Retrieval at the End of the Early Years", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.22, No.12, pp. 1349–1380, 2000. IEEE Computer Society, Washington DC, USA

[2] V.N. Gudivada, "Content-based image retrieval systems (panel)", CSC '95: Proceedings of the 1995 ACM 23rd annual conference on Computer science, pp. 274–280

[3] M.S. Lew, N. Sebe and J.P. Eakins, “Challenges of Image and Video Retrieval”, CIVR ’02: Proceedings of the International Conference on Image and Video Retrieval, pp. 1–6

[4] W. Junqui, R. Cipolla and Z. Hongbin, “Vision-based Global Localization Using a Visual Vocabulary”, ICRA ’05: Proceedings of the 2005 IEEE International Conference on Robotics and Automation, pp. 4230-4235

[5] D.G. Lowe, “Object recognition from local scale-invariant features”, ICCV ’99: Proceedings of the 1999 International Conference on Computer Vision, Vol. 2, pp. 1150–1157

[6] A.P. Dempster, N.M. Laird and D.B. Rubin, “Maximum Likelihood from incomplete data via the EM algorithm”, Journal of the Royal Statistical Society *B*, Vol. 39, pp. 1–38, 1977

[7] T. Hofmann, “Probabilistic latent semantic indexing”, SIGIR ’99: Proceedings of the Conference on Research and Development in Infomration Retrieval, pp. 50–57

[8] A. Bosch, A. Zisserman and X. Muoz, “Scene Classification Via pLSA”, ECCV ’06: Proceedings of European Conference on Computer Vision 2006, Vol. 4, pp. 517–530

[9] K. Fukunaga, “Statistical Pattern Recognition”, Second ed., Academic Press, 1990

[10] D. Comaniciu and P. Meer, “Mean Shift: A Robust Approach Toward Feature Space Analysis”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 5, pp. 603–619, 2002, IEEE Computer Society, Washington DC, USA

[11] R.O. Duda, P.E. Hart and D.G. Stork, “Pattern Classification”, John Wiley and Sons, 2001

[12] K. Mikolajczyk and C. Schmid, “An Affine Invariant Interest Point Detector”, ECCV ’02: Proceedings of European Conference on Computer Vision 2002, Vol. 1, pp. 128–142

[13] B. Schlkopf and A. Smola, “Learning with Kernels”, MIT Press, 2002

[14] C-W. Hsu and C-J. Lin, “A comparison of methods for multi-class support vector machines”, IEEE Transactions on Neural Networks, No. 13, pp. 415–425, 2002, IEEE Computer Society, Washington DC, USA

[15] C-C. Chang and C-J. Lin, “LIBSVM: a library for support vector machines”, 2001.