Abstract

Labeling the semantic content of images is a problem known as image annotation. Automatic image annotation is the process by which a computer system automatically assigns metadata, in the form of captions or keywords, to a digital image. Automatic image annotation often uses computer vision techniques to extract meaningful cues from the image content. It can be applied in image retrieval systems to organize and locate images of interest from a database.

Most of the image database systems employ manual annotation, that is users enter some descriptive keywords when the images are loaded, registered or browsed. Although manual annotation of image content is considered a “best case” in terms of accuracy (meaningful keywords are selected by the human) it is a time consuming and intensive process. In addition, human annotation is subjective: human observers tend to disregard background elements favoring subjects related on their life like humans or animals.

Automatic image annotation can be regarded as multi-class image classification problem with a very large number of classes. Typically, image analysis in the form of extracted feature vectors and the training annotation words, are used by machine learning to automatically apply annotations to new images.

The objective of this thesis is the development of an architecture of classifiers for automatic image annotation. The proposed framework is based on the idea that region-level analysis can improve image scene analysis and content description. The approach based on regions should provide a more robust classification since it is a good compromise between local and global approaches. The building blocks of my thesis will be:

1. Representing the image content by first segmenting the image and then computing meaningful descriptions for each region.

2. Studying data-driven feature selection for reducing the description dimension, devising unsupervised or partially supervised learning strategies to organize the data in homogeneous clusters

3. Designing an architecture of classifiers able to automatically assign a set of tags to a given image.
# Contents

1 Introduction ............................................ 3

2 Statistics of Pictorial Images ......................... 4

3 Image Segmentation .................................... 6
   3.1 The Gestalt Influence to Segmentation .............. 7
   3.2 Segmentation Approaches ........................... 7
   3.3 Histogram Thresholding ................................ 8
   3.4 Clustering ........................................... 9
      3.4.1 K-means ........................................ 9
      3.4.2 Mean Shift ...................................... 10
      3.4.3 Spectral Clustering ............................ 11
      3.4.4 Fast Multiscale Image Segmentation .......... 11

4 Segment descriptors .................................... 12
   4.1 Color Features ...................................... 12
   4.2 Texture Features .................................... 13
   4.3 Geometrical Features ................................ 14

5 Semi-supervised and Unsupervised Learning .......... 14

6 Framework of the project ............................... 15
   6.1 Preprocessing ....................................... 16
   6.2 Representation of an Image Segment ............... 17
   6.3 Architecture ........................................ 17

7 Objectives of the Thesis ................................ 17
   7.1 Objectives ........................................... 17
      7.1.1 Short Term Objective ........................... 18
      7.1.2 Long Term Objective ............................ 18
   7.2 Possible structure of the thesis ................... 18

References ................................................. 19
1 Introduction

Automated annotation of digital pictures has been a highly challenging problem for computer scientists since the invention of computers. The capability of annotating pictures by computers can lead to breakthroughs in a wide range of applications including Web image search, online picture-sharing communities, and scientific experiments.

Image search provided by major search engines such as Google, MSN, and Yahoo! relies on textual descriptions of images found on the Web pages containing the images and the file names of the images. These search engines do not analyze the pixel content of images and hence cannot be used to search unannotated image collections. Although owners of digital images can be requested to provide some descriptive words when depositing the images, the annotation tends to be highly subjective, as mentioned above. A computerized system that accurately suggests annotation tags to users can be very useful. However, automatic annotation of images with a large number of concepts is extremely challenging, a major reason that real-world applications have not appeared.

Humans use a large amount of background and contest knowledge when interpreting an image. With the endowed capability of imagination, we can often see what is not captured in the image itself. For instance, when we look at the pictures in Figure 1, we know that the three images contain branches of a tree and sky on the background, sea waves, and a cloudy sky. In the first image we really notice and understand the presence of the sky thanks to the leaves on the foreground. Similarly, in the last image we recognize the sky because of the clouds: a detail of a clear sky would be much more ambiguous. In the case of the sea the presence of waves is essential to understand its essence; this information, though, is somewhat ambiguous, and it is not enough to discriminate between the sea surface or a lake surface on a windy day. Therefore, we use not only our prior knowledge but also contest informations. We can perceive the meaning of the whole scene, by observing a few details, even if the scene is observed by a close distance or if the view is very partial. With no doubts, it is a very difficult task to empower computers with the capability of mimicking all these levels of human intelligence. However, we can potentially train computers by examples to recognize certain objects and concepts.

The main objective of my thesis is to devise image description methods and classification strategies to build data driven systems for automatic or semi-automatic image annotation. Because of the complexity of this problem, we will consider a restricted family of images, focusing on daily, outdoor images. This approach is not restrictive, since we will focus on data driven solutions (therefore with a different dataset we could analyse other image families).

At first, our approach is based on partitioning an image into multiple regions: homogeneous regions give us guarantee to have approximately a more robust description for a class of “blobs”. Probably, regions larger than a threshold will be used and, certainly, will be described by a set of features. Hence, the next problem will be to find a representation for each segment. The representation should incorporate as much as possible all information that will be useful to discriminate the segment’s content from other contents. The role of chosen features is fundamental since the obtained clusters will depend on such information. At this stage we will need to study inside out various types of descriptors. The architecture of the classifiers will depend on these clusters, for this reason it will be data-driven and will be based on by semi-supervised and unsupervised techniques. We
aim at obtaining a structure, alternative to the one inspired by the natural language (or the human perception), inspired to the descriptions used.

This thesis proposal is organized as follow: Section 2 explores the statistical properties of natural images per scene category and Section 3 introduces the state of the art of segmentation methods and the connected descriptors. Section 4 summarizes the semisupervised techniques focusing on Laplacian Regularized Least Squares and Laplacian Support Vector Machine; Section 5 sketches the possible architecture of our project.

2 Statistics of Pictorial Images

In this section I will show the statistical properties of pictorial images belonging to different categories and their relevance for scene and object categorization tasks. We discuss how second-order statistics are correlated with image categories, scene scale and objects.

The pictorial images are structured according to a hierarchical frame where each basic-level category (at the top of the hierarchy) is most differentiated from one other. The basic categories are the first categories we learn and it has been shown that objects of the same basic-level category share similar common features [1]. In the case of object categories like faces for instance, the common feature is given by the regularity of the distribution of pixel intensities. Instead, for pictorial images, as Oliva and Torralba [2] pointed out, the common features can be obtained observing the following three factors:

1. There is a correlation between an object and its background.

2. The background exhibits texture and colour pattern that are common to each image category.

3. The scale factor (viewpoint of the observer) plays a key role in the change of statistics.
Table 1: In the two rows are respectively shown images of man-made scenes and natural scenes

The first assumption can be immediately noticed observing the average of images constrained to have a particular object at fixed scale. There are common patterns when we found specific objects in the images.

They also demonstrated these asserts showing the correlation of second-order-statistics with image categories, scene scale and objects. Statitics of these images follow particular regularities: it has been observed that there is a bias in the distribution of the orientation of the average power spectrum of natural images. On first, the vertical and horizontal orientations are more frequent than obliques and the energy of the spectral signature of the man-made scene is very different by the natural scene. In images of man-made scenes vertical and horizontal orientations are quite well-marked and the spectrum lies usually on low-frequencies. The spectral signatures of natural enviroments have a broader variation in spectral shapes and there is not a marked orientation for things closely observed. In fact, another aspect concerns the dominant spatial orientations for varing scales: typically, the spectral signature of pictures of large-scale scenes are dominant by horyzon while scene closer to the observer become isotropic. Hence, the connection between structure of the spectral signature and scale scene are mainly two:

1. When the scale increases the viewpoints of the observer become limited and predictable. Therefore close-up views on natural structures have a tendency to be isotropic in orientations while the far-views become more constrained.

2. The parts that compose one scene differ strongly from one scale to another scale, due to functional constraints and to the physical processes that shape the space at each scale. For instance, in the case of the images obtained with a wide viewing angle, different statistics will characterize the top and the bottom half of the image (for instance, smooth texture of the sky on top, cluttered patterns on bottom, etc.).

The relationship between scale, levels of detail, and categories is a strong motivation for our work. Starting from this analysis carried out on whole images, we will investigate the importance of scene statistics on homogeneous regions of the image. This will allow us to extend this analysis to a wider range of prototype classes.
Figure 2: Two mean power spectra averaged from 100 images of the man-made scenes and natural environments. We can notice that, in the first case, horizontal and vertical orientations are more evident.

3 Image Segmentation

Segmentation is the partitioning of a digital image into multiple regions (sets of pixels), according to a given criterion. The goal of segmentation is typically to locate object of interest, and, for this reason, segmentation is often related to object detection problems.

Since a prototype partitioning of an image does not exist, every segmentation algorithm fulfills the requirements depending on its application field. Therefore segmentation methods are very specific and may be very different. Having said so, the partitioning of the image should satisfy the following properties:

- Distinct: no pixel belong to different regions.
- Complete: all pixels are assigned to at least one region
- Connected: all pixels belonging to a region are connected
- Homogeneous: all regions are homogeneous with respect to a given criterion.

The segmentation problem can be formally written as follow. Given an image $I_{rc} = \{x_{ij} : i = 1, \ldots, r, \ j = 1, \ldots, c, \ x_{ij} \in S\}$ and a similarity predicate $H$ defined for each set of pixels, a region $R_i$ is defined as $\{(i_{l1}, j_{l1}), (i_{l2}, j_{l2}), \ldots, (i_{ln}, j_{ln})\} \subseteq \{1, \ldots, r\} \times \{1, \ldots, c\}$

Then, the segmentation of $I_{rc}$ is a partition $P = R_1, R_2, \ldots, R_K$ such that:

- $R_i \cap R_j = \phi$ \quad $i, j = 1, \ldots, k$ \quad $i \neq j$ (Distinct)
- $\bigcup_{i=1}^{K} R_i = \{1, \ldots, r\} \times \{1, \ldots, c\}$ (Complete)
- $R_i$ is “connected” (Connected)
- $H(R_i) = \text{true} \ \forall R_i$ (Homogeneous) $H(R_i \cap R_j) = \text{false} \ \forall R_i, R_j$
3.1 The Gestalt Influence to Segmentation

At the beginning of the last century Gestalt Psychology [3] attached the importance to the form-forming capability of our sense, particularly with respect to the visual recognition of figures and whole forms rather than just a collection of simple lines and curves. Therefore several properties, such as similarity or proximity were listed as key factors for visual grouping. In the case of image segmentation this is evident, in fact many segmentation strategies have reference to Gestalt Principles. The Gestalt Principles of Visual Organization are the followings:

- **Figure and Background**: the relationship between figure and background allow to read image. The dominant elements are perceived as figure while the residue as background. The Tresholding method has reference to this principle.

- **Proximity**: human perception tends to group features that are close to each other. Clustering, region growing and region merging use this idea to partition the image.

- **Closure**: missing information is inferred, in fact the human eye tends to fill empty spaces and complete no-end shapes. Morphological methos and edge detection have a reference to this concept.

- **Continuation**: the human attention is catched by the perpetuation of a line (rather than a torsion).

- **Similarity**: similar visual elements are grouped with respect to their shape, size color or orientation. Clustering, region growing, region merging and statistical methods use this idea to partition the image.

- **Common destiny**: objects moving along the same direction are perceived as the same identity.

- **Parallelism**: parallel lines tend to group together.

- **Common region**: figures in the same end region are clustered together.

- **Symmetry**: the human perception tends to assemble things constituting a symmetry rather than look at single elements.

3.2 Segmentation Approaches

The literature on segmentation is very rich. In this overview I follow a taxonomy suggested in [4]. They divide the segmentation algorithms into:

1. **Feature space-based techniques** (also known as pixel-based techniques): based on the idea that, for different objects, pixels could manifest themselves as cluster or peaks into histograms. These two approaches share a common property: they work in a certain feature space, which may be one of the color spaces or a space induced by other attributes, and they generally neglect the spatial relationships. For this reason, the denomination is feature space-based techniques and they are separately considered by the image-domain techniques.
2. **Image-domain based techniques**: the feature space-based techniques do not guarantee that regions also show spatial compactness, which is a second desirable property in segmentation applications beside homogeneity. In fact, cluster analysis and histogram thresholding do not account for the spatial locations of pixels; the description they provide is global and it does not exploit the important fact that points of a same object are usually spatially close due to surface coherence. Usually, the available techniques can be divide into three main groups: *Split and Merge*, *Region Growing* and *Edge based techniques*. The first method is iterative: it starts with an initial inhomogeneous partition of the image (usually the initial segment is the image itself) and they keep performing splitting and merging until homogeneous partitions with respect to a certain statistic criterion are obtained. In the second approach an homogeneous region of an image may be obtained through a growth process which, starting from a preselected seed, progressively agglomerates points around it satisfying a certain homogeneity criterion. Finally, Edge based techniques provide the segmentation by detecting the edges among regions. The central challenge to edge finding techniques is to find procedures that produce closed contours around the objects of interest.

3. **Physics based techniques**: the algorithms examined so far are certainly prone to segmentation errors if the objects portrayed in the images are affected by highlights shadowing and shadows. These phenomena cause the appearance of grey-levels or color change, more or less drastically, on uniformly surfaces. There algorithms above mentioned are very likely to return oversegmented regions. The only way to overcome this drawback is to analyze how light interact with materials and to introduce models of this physical interaction in the segmentation algorithms.

In the following sections we will report more details on a selection of methods.

### 3.3 Histogram Thresholding

Histogram Thresholding is among the most popular techniques for segmenting grey-level images. This is primarily due to the fact that is an easy efficient method to implement and provides satisfactory results in many cases. In various applications, it can be effectively used as the initial step in more sophisticated image analysis tasks. Examples of such application include segmentation of brain tissue and/or tumors in magnetic resonance (MR) images and quantification of nuclei of cells and chromosome in microscope images. The basic method is trivial: the peaks and the valleys of the 1D brightness histogram can be easily identified, respectively, with objects and backgrounds of the grey-level images. The threshold for splitting the image, can be arbitrarily fixed or chosen with respect to a given, usually statistic, automatic criterion. The most famous approaches are the *peaks and valleys method* and the *global threshold selection*. The first one founds the two higher peaks and afterwards founds the deepest valley between them. The second one is an iterative method: one can choose an initial threshold, calculate a segmentation and afterwards compute a new threshold by the means of segmentation until point of convergence. Unfortunately, signal and background peaks are usually not so ideally separated (due to presence of noise or poor contrast), and the choice of the threshold is problematic. In fact, typical histogram to analyze in these cases are often bimodal but with peaks not
well separated. Besides, if the desired goal is to partitioning the images in more than two parts, these approaches can be repeated recursively. Although these techniques are not so accurate, approaches that consider colour images, either individual histograms of red, green and blue channels or a 3-D (2D) histogram were produced.

3.4 Clustering

Clustering can be broadly defined as a nonsupervised classification of objects in which one has to generate classes or partitions without any a priori knowledge [1]. The clustering problem can be stated as follows: let us suppose that we have \( M \) patterns \( x_1, \ldots, x_M \) within a certain pattern space \( S \). The process of clustering consists in determining the regions \( S_1, \ldots, S_K \) such that every \( x_m, \ m = 1, \ldots, M \), belongs to one of these regions and no \( x_m \) belongs to two regions at the same time, i.e., \( \bigcup_{k=1}^{K} S_k = S \) and \( S_i \cap S_j = \emptyset \). The classification of patterns into classes follows the general common sense principle that objects within each class should show a high degree of similarity while across different classes they should exhibit very low affinity. In the following sections I will explain the techniques that we implemented and used.

3.4.1 K-means

A classical technique for image segmentation is the k-means (or c-means) algorithm [5], widely adopted also for vector quantization and data compression. The method belong to the class of partional clustering: it consists of dividing an image without any hierarchal structure on segments.

This algorithm aims at minimizing an objective function, usually a squared error function:

\[
J = \sum_{j=1}^{K} \sum_{i=1}^{n} \| x_i^{(j)} - c_j \|^2
\]

where \( \| x_i^{(j)} - c_j \| \) is a chosen distance measure between a data point \( x_i^{(j)} \) and the cluster centre \( c_j \) and \( k \) is the number of clusters and. K-means is an iterative algorithm, organized in the following steps:

1. Initialize the algorithm choosing randomly \( K \) points in the feature space: these points represent the initial centroids of the cluster
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the \( K \) centroids.
4. Repeat Steps b and c until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centres and depends on the number \( k \) of chosen clusters (or segments).
3.4.2 Mean Shift

Comaniciu and Meer [6] [7] resort instead to the mean shift algorithm which is a non-parametric procedure for estimating density gradients of pattern distributions. The method they developed is conceptually very simple. It is based on the idea of iteratively shifting a fixed size window by averaging the data points at each step. The kernel density estimator is the following:

\[
\hat{f}_{h,K}(x) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^{n} k\left(\frac{\|x - x_i\|}{h}\right)
\]

where \( n \) is the data number, \( K \) is a multivariate Kernel (\( k \) is its profile), \( h \) is the bandwidth, \( d \) is data dimension and \( c_{k,d} \) is a constant depending on \( K \) and \( d \). The last parameter is important because controls the under-over smoothing of the density estimate. They used two kernels: Gaussian and the Epanechnikov that minimize Asymptotic Mean Integrated Square Error. The modes are located among the zeros of the gradient and the mean shift procedure is an elegant way to locate this zeros without estimating the density. Using a differentiable kernel the density gradient can be written as follow:

\[
\hat{\nabla}f_{h,K}(x) = \frac{2c_{k,d}}{nh^{d+2}} \left[ \sum_{i=1}^{n} g\left(\frac{\|x - x_i\|}{h}\right) \right] \left[ \frac{\sum_{i=1}^{n} x_i g\left(\frac{\|x - x_i\|}{h}\right)}{\sum_{i=1}^{n} g\left(\|x - x_i\|^2\right)} - x \right]
\]

were \( g = -k'(x) \). The first term is proportional to the density estimate at \( x \) computed with the kernel \( G \) and the second term is the mean-shift vector. They showed that the mean-shift vector points toward the direction of the maximum increase in the density:

\[
m_{h,G}(x) = \frac{1}{2}h^2 c \frac{\hat{\nabla}f_{h,K}(x)}{\hat{f}_{h,G}(x)}
\]

Given this relationship, showing that the local mean is shifted toward the region in which the majority of the points reside, they provided the algorithm to filter and segment the images.

Because of mean shift vector is aligned with local gradient, the algorithm is an iterative procedure that searches stationary points of the density:

- Computation of the mean-shift vector
- Translation of the kernel (window) \( G(x) \) by \( m_{h,G} \)

Filtering is a preprocessing step for the segmentation: after selecting the pixel features (they choose points in the joint domain spatial-range) and reaching the stationary points with the iterative procedure (filtering) it needs to group pixels in the joint domain and eliminate spatial region containing less of \( M \) pixels. This method converges and does not require any prior knowledge on the number of clusters as k-means does. Besides, it is fast even in the presence of many data and therefore suitable for image segmentation.
3.4.3 Spectral Clustering

Another well established approach was proposed by Shi and Malik [8]. They tackle image segmentation via clustering as a graph partitioning problem. They represent the set of points in an arbitrary feature space as a weighted undirected graph \( G = (V, E) \), where the nodes are the points in the chosen feature space and edges are established between each pair of nodes. The weight \( w(i, j) \) of each edge is a function of the similarity between nodes \( i \) and \( j \). The goal is to partition the set of vertices \( V \) into disjoint sets \( V_1, \ldots, V_m \) such that a predefined similarity measure is high for vertices within the same set and low across different sets. Therefore, they proposed to minimize the following measure also named Normalized Cut:

\[
Ncut(A, B) = \frac{\text{cut}(A, B)}{\text{assoc}(A, V)} + \frac{\text{cut}(B, A)}{\text{assoc}(B, A)}
\]

where \( \text{cut}(A, B) \) is the sum of the weights between \( A \) and \( B \) and \( \text{assoc}(A, V) \) is the total connection from nodes in to all nodes in the graph. Unfortunately, minimize \( Ncut \) is NP-hard, so an approximate discrete solution can be found by solving the minimum of the following function:

\[
\min_x x^T (D - W) x \quad \text{with the constrain } x^T D 1 = 0,
\]

where \( x \) is the indicator vector (representing \( A \) and \( V-A \) sets), \( D \) is the diagonal matrix of the total connection with a node and \( W \) is the weight matrix of the graph. Therefore, the method to segment an image consist of to recursively bipartite an weighted Graph where \( V=\text{pixels} \) and \( E=\text{similarity between the pixels} \).

Although this method prevents the over-segmentation of the image, it works in \( O(n^3) \) and require an computational space in \( O(n^2) \) where \( n \) denotes the number of pixels in the image. Besides these technique does not exploit a hierarchical or multiscale approach that could be interesting about conclusions in section 2.

3.4.4 Fast Multiscale Image Segmentation

For the reason illustrated above I explored an algorithm provides a hierarchical decomposition of the image into segments [9][10][11]. The algorithm detects the segments by applying a process of recursive coarsening of a Graph in which the same minimization problem above-mentioned is represented with fewer and fewer variables producing an irregular pyramid. Once the pyramid is completed it is scanned from the top down to associate pixels close to the boundaries of segments with the appropriate segment. The procedure can be summarizes into two phases:

1. **Detecting the Salient Segments.** The initial nodes of the graph are image pixels and the couplings values (matrix \( W \)) are set with respect to an “edgeness” measure. During the scanning of pixels they agglomerate nodes strongly connected, otherwise they create new seeds in the list. Then, they recompute weights among new nodes of the coarsen graph and determine the salient segments with a coarsening rule. It will need calculate the Normalized Cuts for a coarse graph and renavigate the pyramid from the top to have an approximation of the \( Ncut \).
2. **Sharpening Boundaries**: in this phase they give a rule to choose from the new weights the membership to a segment. They fixed parameters to decide which pixels on boundary (because the new graph contain “fuzzy” values) belong to a segments.

4 Segment descriptors

Image segmentation allows us to identify homogeneous regions of an image. The next step of our analysis will consist in trying to group similar regions and, if possible, discriminate among regions containing different information. For this reason we need to represent region information in a uniform way, so that we can compare and classify different regions. A region description can be though of as a list of features, carrying information about the region content. Common features are related to color, texture or shape information.

4.1 Color Features

Color information has been widely used both in image segmentation and classification. The reasons are many: color descriptions are easy to implement, intuitive to use and in many simple problems they are discriminant enough. They are usually robust to the presence of noise, changes in resolution, orientation, resizing, and translation. Combined with image segmentation, color features can be used to describe the appearance of the image and even generate semantic annotations.

A typical description for color information is the color histogram, that describes the color distribution of an image.

Color histograms (or histograms in general) can be compared using one of the many histograms similarity or distance measures proposed in the literature. Here we mention a few of them.

* Minkosky-Form distance

\[ d_{L_1}(I, H) = \| I - H \|_{L_1} = \left( \sum_{i=1}^{l=n} |i_l - h_l|^r \right)^{\frac{1}{r}} \]

The $L_1$ distance is often used for computing dissimilarity between color images [12], but other common usages are $L_2$ and $L_\infty$.

* Histogram Intersection [12]

\[ d_\cap(I, H) = 1 - \frac{\sum_i \min(h_i, k_i)}{\sum_i k_i} \]

to matching histogram with different areas. Other possible Bin-By-Bin metrics include *Kullback-Leibler Divergence and Jeffrey Divergence* and $\chi^2$ *Statistics* [13][14].

There are also techniques comparing non corresponding-bins that make use of the *ground distance* $d_{ij}$, defined as the distance between the representative features for bin $i$ and bin $j$. Among them we can found the *Quadratic-form distance*, *Match distance* and *Kolmogorov-Smirnov distance*.

However, for some images often only small fraction of the bins contain significant information, while most others are hardly populated. A finely quantized histogram is
highly inefficient in this case. In brief, because histograms are fixed-size structures, they cannot achieve a good balance between expressiveness and efficiency. On the other hand, there is another method trying to remedy to this problem to represent color feature: color signature. It is based on the idea that a histogram \( \{h_i\} \) can be view as a signature in which vectors \( i \) index a set of clusters defined by a fixed a priori partitioning of the underlying space. In this sense, a useful metric between features, used in the past for image retrieval, is the Earth Mover's Distance [13].

Both color histograms and signatures discard all the correlation information between different pixels. The image spatial support is destroyed and only frequency information is considered. Alternatively one could choose second-order statistics descriptions, such as color correlograms, that consider at what frequencies couples of colors co-occur at a given distance in the reference image. Color correlograms carry a richer information but they are even sparser than histograms, for this reason they should be used only in combination with appropriate space optimization strategies.

### 4.2 Texture Features

Texture is a very important visual cue for identifying regions with homogeneous periodic patterns like weaves, vegetation, walls, material and every other textured surfaces. It further has been shown that texture features have a very high discriminanting power, enabling distinction that would otherwise prove impossible. An important first step to identify the perceived qualities of texture is to build mathematical models representing the intensity variations in an image. Modelling this physical variations is a problematic issue because it is not clear what one mean with texture. However, texture has a number of properties which are generally assumed to be true [15]:

- Texture is a property area; the texture point is undefined
- Texture involves the spatial distribution of the gray level or colors.
- Texture in an image can be perceived at different scales or levels of resolution
- A region is perceived to have texture when the number of primitive objects in the region is large

There are many methods to extract qualities of texture to explore in my thesis belonging to the following classes:

- **Statistical methods**: based on the spatial distribution of gray values. The most famous approach are Co-occurrence Matrices and Autocorrelation Features.

- **Geometrical methods**: depends upon the geometric properties of the “texture elements” (objects composing a texture). Once the texture elements are identified in the image, there are two major approach to analyzing the texture. One computes statistical properties from the extracted texture elements and utilizes these as features (Voronoi Tessellation-Moments). The other one tries to extract the placement rule that describe the texture (Structural Methods).
• **Model based methods**: based on the construction of an image model can be used not only to synthesize but also to describe texture (*Random Field Models, Fractals*).

• **Signal Processing methods**: depends on the frequence analysis of the image (according to the behavior of the human brain). Texture is especially suited for this type of analysis because of its properties. The most famous approach are:

  1. Spatial Domain Filters
  2. Fourier Domain Filtering
  3. Gabor and Wavelets models

These technique, trying to compute certain features from filtered images, are very suitable for segmentation and classification tasks; for this reason could be better explored in my thesis.

### 4.3 Geometrical Features

Geometric features encompass position, size and shape information. All those characteristics can be used to further refine a classification obtained through color and texture features. Hence, for instance, the sky regions could have most benefit by the feature “position” and by “area” respect to the entire image. Not only characteristics of the whole image can be considered. Other features like lines, curves, ellipses or general closed contours could provides a large number of additional information to the region to analyze. In this sense there are many techniques allowing to extract features by an image segment [16]

### 5 Semi-supervised and Unsupervised Learning

From an engineering standpoint, it is clear that collecting labeled data is generally much more cumbersome than collecting unlabeled data. This is particularly true if the aim is to collect a multitude of images that represents a significant number of classes. In fact, our purpose is to collect a training set of regions obtained by segmentation representing classes belonging to pictorial images. Hence, an approach making better use of unlabeled data should improve recognition and decrease the need of labelling a large number of data. Actually, most natural learning occurs in the semisupervised regime: a large amount of unlabeled data are collected by humans to extract information that is useful for generalization.

In the *Semi-supervised Learning* paradigm, the goal is to understand how the natural learning comes about and which factors play a key role. The main idea is to exploit the geometry of the unlabeled data. Such an information is entailed in the marginal distribution, underlying the unlabeled data, which is fixed but unknown and can be faithfully approximated if enough unlabeled data are available. The main idea of semi-supervised techniques is to incorporate the geometric information to extend the more established framework of supervised learning [17]. We now specialize a bit more the above discussion.
The standard framework of learning from examples [17] is formally defined as follows: we are given a training set of labeled examples \((x_i, y_i), i = 1, \ldots, n\), generated according to fixed (but unknown) distribution \(P(x, y)\), and the goal is to find a function \(f\) such that \(f(x_{new}) \sim y_{new}\). The goal is not to describe the training data but to model the input-output relation they are sampled from. To achieve such a goal the main intuition is that we have to avoid overfitting the data. A large class of algorithms, namely regularization networks [18], can be written as

\[
f* = \arg\min_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i, f) + \lambda_A \|f\|_K^2.
\]

The function \(\ell(\cdot, \cdot, \cdot)\) is the loss function measuring the error on the data, different algorithms are implemented by choosing different loss: the square loss \((y_i - f(x_i))^2\) gives Regularized Least Squares and the hinge-loss \(\max[0, 1 - y_i f(x_i)]\) gives Support Vector Machines. The term \(\|f\|_K\) is a norm in a reproducing kernel Hilbert space [19] which is a penalization imposing smoothness conditions on the class of possible solutions.

Following [20] if we are in semi-supervised setting we can add an extra penalization term \(I\|f\|_K^2\) encoding information on the intrinsic geometry of the data. For example we would like to choose \(\|f\|_K^2 = L\), where \(L\) is the laplacian operator, but such a choice requires the knowledge of the marginal measure \(P_X(x)\) describing the input points. The idea is that such a penalization term can be faithfully approximated if we have enough unlabeled data \(x_1, \ldots, x_u\) sampled according to \(P_X(x)\). Thus, given a set of \(l\) labeled examples \(\{(x_i, y_i)\}_{i=1}^{l}\) and a set of \(u\) unlabeled examples \(\{x_j\}_{j=l+1}^{l+u}\) we can consider the following optimization problem:

\[
f* = \arg\min_{f \in \mathcal{H}_k} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \lambda_A \|f\|_K^2 + \frac{\lambda_I}{(u+l)^2} f^T L f^T
\]

where \(f = [f(x_1), f(x_2), \ldots, f(x_n)]^T\) and \(L\) is the the Graph Laplacian given by \(D - W\). Interestingly, the minimizer of the above problem has this form:

\[
f^*(x) = \sum_{i=1}^{l+u} \alpha_i K(x_i, x),
\]

where \(K(x_i, x)\) stands for the kernel and \(\alpha_i\) also contain a term depending on unlabeled data. When \(\lambda_I = 0\) the solution is the same for standard supervised algorithms, while if we have unlabeled points we can improve the results. We note that usually there is no condition about the relation between marginal and conditional distribution but while using the above algorithm we are implicitly assuming that if two points \(x_1, x_2\) are close in the intrinsic geometry then the conditional distributions \(P(y|x_1)\) and \(P(y|x_2)\) are similar, namely the manifold assumption.

### 6 Framework of the project

The main objective of this project is the development of an architecture of classifiers for automatic image annotation. This architecture will be the core of a system that reads
an image, extracts meaningful segments with a segmentation procedure, and assigns a label to each of the segments that appears similar to one of the classes modeled by the architecture. The ambition is to build an architecture that can be expanded easily if we need to add new classes. The aim is to detect and recognize homogeneous background areas within an image.

We will concentrate on a subset of all the possible pictorial images which includes daily, outdoor images. This choice is not restrictive, as it has shown in the past, it is possible to classify automatically images belonging to these classes [21]. We will consider a set of classes of interest that include: sea, sky, grass, vegetation (trees and other vegetation observed from a distance), rocks (mountains, reefs), snow, urban pavement, country roads and paths, sand, pebble beaches, cityscapes observed from a distance and crowds from a distance.

Some of these classes can be further sub-classified, such as sea (calm, wavy, rough, ..), sky (clear, partially clouded, clouded, ..). It will also consider classes that are something between background features and objects with a well defined shape:

1. Buildings and monuments observed from a relatively close distance, so that they can be seen as background elements (but at the same time they are characterized by a well defined geometry);

2. Natural elements observed from close range (mountains, rivers).

6.1 Preprocessing

The preprocessing stage aims at identifying homogeneous parts of the image (image segments), and at designing methods for segments description.

To this purpose it may be useful to analyse how the segments vary while varying the image scale. Ideally we would like to identify segments at their most representative scale (minimizing oversegmentation and undersegmentation of areas with a complex texture structure, such as cityscapes or pebble beaches). As initial step, I could use an standard segmentation approach, as we see on section 3. I focused on clustering approach that could be reuse in a second phase of my thesis to drive labelling of the data.

It is not clear if we need a multiscale approach to segmentation or if it will be enough to incorporate multiscale segments descriptors, that are quite common among the texture descriptors. So, the future plan is to investigate the list of following techniques:

- Multiscale segmentation
- Feature extraction (color histograms or color signatures, texture descriptions, shape descriptors, see the methods above mentioned 3)
- Multiscale features
- Multiscale image descriptions, such as wavelets
6.2 Representation of an Image Segment

Let us assume that an image has been segmented in a number of regions, that are homogeneous with respect to some descriptor.

The next problem is to find a representation for each segment. The representation should incorporate as much as possible all information that will be useful to discriminate the segment’s content from other contents.

To build a system that can be easily adapted to new classes we will look for a global code-book or vocabulary that describes a generic segment. A possible code-book will contain color information, texture descriptions, shape descriptors. Maybe also local features to describe possible elements of interest. It may be useful to describe the segment position within the original image relative to the image size. It is not clear how to include scale information, i.e., it is not clear if the description should contain information about the segment at different scales.

Probably a feature selection step will follow. In this phase we will identify a number of features belonging to the global code-book that are characteristic of a given class. Notice that a feature may be representative of a class either because it often appears, with a certain value, in example segments of the class of interest (and possible does not appear in examples belonging to other classes), or because it never appears in a class (e.g., red shades for the sea classes).

6.3 Architecture

The main issue from the architecture point of view is to design a structure of classifiers, possibly with a hierarchic structure. As a starting point we will consider the following two elements:

1. There is no obvious relationship between the semantics of a class as it is given by a human observer and the one induced by the image descriptors that we will use.

2. The possible links between multiscale image descriptions and the hierarchy of classifiers. The interclass distance may change if the images are observed at a different resolution, for instance at a low resolution similar classes (of sub-classes of the same parent class) may overlap.

These two elements suggest that we investigate data driven strategies to infer elements about the architecture structure. In particular we will investigate the use of unsupervised or partially labeled techniques to understand possible subclasses of a given class.

7 Objectives of the Thesis

In this section I outline the objectives of the next two years, describing both short term and long term objectives. I will also include a tentative structure of my thesis.

7.1 Objectives

Describing the content of an image is a problem which can be seen from many different points of view: recognition, classification, detection or image retrieval give a different
interpretation of what is represented in an image. The main objective of this thesis is approaching automatic image annotation by segmentation, region description, feature selection and clustering. One of the challenges of my work, in addition to understand the image content, is to see how unsupervised techniques can help us to obtain a data-driven procedure for the design of the classifiers architecture. For instance, if we only use color it is possible that some classes, even if they are semantically different, will merge (for instance clear sky and calm sea). For these reason we can say that a new and interesting challenge is using clustering of similar feature to create a architecture for classification. Therefore the upcoming step is to use trivial feature, as color histogram of regions, observing the results. the following scheme resumes the Objectives of my thesis’ work:

7.1.1 Short Term Objective

Short terms objectives of my thesis refer to the next 6/8 months: as before mentioned, a part of the work will be mainly devoted to study state of the art methods and to implement or acquire code concerning the most promising techniques to feature description and multiscale segmentation. At the same time the work will consist of:

1. Collecting a relatively small data set including representative classes, such as sky, sea and grass,

2. Choose simple descriptions (color histograms and similar),

3. Apply simple clustering methods at fixed scales and study the distribution of the obtained clusters.

7.1.2 Long Term Objective

Long terms objectives will cover the following 8/12 months: I will investigate and implement a number of classification algorithms, including the recently proposed hierarchical classification models (see [22] and references); possibly we will consider learning algorithms that deal with vector labeled outputs (see for instance [23]).

However, the main objective of this phase is to build an architecture of classifiers that describe the hierarchy and the similarity relationships between classes. We aim at obtaining a structure, alternative to the one inspired by the natural language (or the human perception), inspired to the descriptions used.

7.2 Possible structure of the thesis

The structure of the thesis could be divided into three main parts (each of them could be arranged in one or more chapters):

1. Introduction: introduction of the thesis with respect to the research area, describing motivation, contribution and importance to the work done.

2. Segmentation and Region descriptors: an overview: a review of different segmentation approaches and region descriptors like color, shape and texture and evaluations with respect to clustering.
3. *Regions classification: the architecture:* an existing overview of the classifications methods used in this thesis for annotating the image. The structure of the project: the exploited segmentation methods, the considered features and the results.

**References**


