A MEDICAL IMAGE RETRIEVAL SYSTEM BASED ON SEMANTIC ANNOTATIONS

Francesco Maiorana
Department of Electrical, Electronics and Informatics
University of Catania
Viale A. Doria, 6, 95125 Catania
Italy
francesco.maiorana@dieei.unict.it

ABSTRACT
This paper presents the design and implementation of a semantic Content Based Image Retrieval Systems (CBIR) developed in Matlab from scratch by choosing a combination of texture, color and shape as low level features to represent the images, and by using a multi-labeling classifier to associate these low level features to a semantic label. We used the Bayes Point machine classifier to classify the images. The classification results are further enhanced by using an explicit relevance feedback algorithm. The system is tested on a set of medical images combined with other types of images and the results are presented.

KEY WORDS
Medical Content Based Image Retrieval, Semantic image annotation, Bayes Point Machine, Relevance feedback.

1. Introduction
An increasing amount of images are being produced especially in the medical domain where new techniques related to 3D produce a fast growing amount of digital images. For these reasons a vast research activity has been done on the field of Content Based Image Retrieval systems (CBIR). CBIR involves knowledge from various discipline with many difficulties or gaps to fill. In [1] the authors identify 14 gaps that are grouped in four categories:

1) Content gaps: This group of gaps addresses the “modelling, understanding, and use of images from the standpoint of a user”. The semantic gap is the most prominent examples of gaps in this category. For semantic gap is meant a missing link between low level numerical features automatically extracted by means of computer algorithms and semantic or high level features assigned by a human domain expert. The semantic meaning can be assigned by means of keyword, free text or ontology concepts to describe the content of an image.
2) Feature Gaps: These gaps correspond to:
   a. the inadequacies of the chosen numerical features to characterize the image content.
   b. the practical difficulties of extracting these features from the images.
3) Performance Gaps: “the lack of evaluation of CBIR system performance and its benefit in health care”.
4) Usability Gaps: describe how easy the use of the system is from the perspective of the end user. In this category the authors mention the Query Gap, the Feedback Gap and the Refinement Gap.

In [2] the authors classify user queries into three types:

1) Retrieval by low level feature such as texture, colour or shape via query by example.
2) Retrieval of images identified by a semantic feature such as “retrieve a sunset”.
3) Retrieve of image by means of abstract semantic attributes such as retrieval of image with emotional message such as “retrieve image of joyful kids”.

Queries belonging to the second and third category are identified as semantic image retrieval; the gap between the first and the second category identifies the semantic gap.

In this paper we try to address the semantic gap of CBIR systems in the medical domain.

Extensive research has been done to reduce the semantic gap and image annotation systems based on image semantic content have been developed. For a review on content based image annotation and retrieval system the reader can reference [3-6].

In [3] the authors review the automatic image annotation systems (AIA) that learn semantic concepts from low level features and use the model to label new images. The review deals with feature extraction and image representation reviewing the major techniques used for:

- image segmentation that can be performed by:
  o dividing the image into fixed size and position blocks.
- Feature extraction:
  - Colour is deemed as one of the most important features for an image: this requires defining the colour space and then extracting colour features that can be based on colour histogram, colour moments, colour coherence vector, colour correlogram, dominant colour descriptor (DCD), colour layout descriptor (CLD), colour structure descriptor (CSD), scalable colour descriptor (SCD).
  - Texture: it is measured from a group of pixels and has a strong discrimination capacity. The methodology to extract texture has been classified into:
    - Spatial: texture is computed considering the pixel statistics in the original image domain. Techniques can be divided into structural, statistical or model based.
    - Spectral: the image is transformed in the spectral domain. Techniques includes Fourier Transform, discrete cosine transform, wavlet and Gabor filters.
  - Shape: is usually classified into contour based that calculates features based on the boundary and region based that extracts features based on the entire region using shape descriptors like area, movements, circularity and eccentricity.
  - Spatial relationship deals with the relation of objects in the image as well as their location.

- Automatic image annotation techniques to reduce the semantic gap. In [4] the authors classify the major techniques used to reduce the semantic gap using low level features to learn semantic concepts:
  - Using ontology to define high level concepts and guide image labelling and querying. In [5] the author reviews the image annotation methods that are classified as: free text, based on keywords and based on ontologies that offer the benefit of structuring keyword in a hierarchy or other kind of relationship making it easier to resolve problems related to ambiguous keywords.
  - Using supervised and unsupervised learning methods. These can be further divided into single labelling and multi-labelling and web based annotations that use image metadata to annotate the image. Among the supervised methods common techniques are based on probabilistic classifiers, Support Vector Machines (SVM), Artificial Neural Networks (ANN), decision trees, and k-Nearest Neighbors. Techniques such as majority voting, bagging, boosting and stacked generalization are used to combine multiple classifiers.
    - Relevance feedback from user to improve the learning model.
    - Semantic template which is a map between high level concept and low level visual feature.
    - Use of both textual image information such as metadata and its visual content.

In [6] the authors review 50 image annotation systems. Among these 21 use block based image features, 15 use object based segmentation; 34 systems assign one keyword, only 8 systems assign multiple keywords, by using region or block based image features; 14 systems use colour features; the majority, 19, us colour and texture; only 6 systems use other features, such as shape; 18 systems use SVM, 7 K-NN, 8 Naïve Bayes, 1 system decision tree and two use template matching; and 33 systems use Corel as ground truth dataset.

Other problem related to AIA as reported in [3] are related to scale increasing factors arousing when the number of labels increase, the lack of a common image database for AIA training and evaluation and the lack of a standard vocabulary and taxonomy for annotation.

The work in [5] reviews some general purpose taxonomy and ontologies. An important step in the direction of image classification has been done in [7] where the authors report an image classification scheme using the IRMA code. The images are analysed into six layers of abstraction such as imaging modality (x-ray, ultrasound, magnetic resonance imaging, optical imaging), body orientation (anterior-posterior or sagittal), anatomical region, biological system and so on.

Ontologies can also be used to better organize and share information on images in an integrated system such as the one proposed, in another domain[8].

In the medical domain a step in this direction has been described in [9] where the authors uses anatomical, disease, and contextual ontologies for image tagging and ontological reasoning for improving image analysis and retrieval.

For image database collection we would like to recall the Casimage project [10] with more than 80,000 images stored, some of them publicly available on Internet. Other initiatives started in the United States [11] or Europe [12] aim to facilitate image sharing.

The Medical Imaging Resource Centre (MIRC) [13] of the Radiological Society of North America (RSNA) allows one to query several databases (including Casimage) by textual queries. Some of the databases available are also used in ImageCLEF [14], an
international competition that allows for comparison of performance of various systems on the same set of images for the same task [15].

Another important collection of images also used in ImageCLED is represented by the Goldminer search engine [16] of the American Roentgen Ray Society (ARRS) that make accessible more than 175,000 images peer-reviewed from many radiology journals. The database at the moment can be searched by text queries on structural data such as age, gender or free text. These image collections can be used for educational purpose by collecting expert annotations and by making these annotations freely available to a community in a shared design memory architecture such as the one proposed in [17][18].

A promising area of research is to integrate large image collection by using grid technologies [19] [20] to securely store and retrieve large collection of images that can take advantage of the computing power of the grid for similarity on line search.

In this work we present an automatic annotation system developed in Matlab based both on colour, texture and shape feature extraction algorithms, that uses a multi-classification combined with majority voting based on a Bayes Point Machine (BPM) classifier for supervised model construction and automatic image labelling with explicit user relevance feedback to adjust the model

This paper is organized as follows: section 2 reviews the techniques used in the CBIR system, section 3 describes the developed systems, section 4 presents some results and section 5 draws some conclusions and highlights future works.

2. Techniques Used In The CBIR System

We developed a CBIR system that is designed along the following major components:

1) Image segmentation
2) Low level features extraction
3) Supervised multiclass classification by means of a Bayes Point Machine (BPM) classifier to associate the low level features to one or more semantic label. For each assigned semantic label a probability of correctness is provided to the user so it is possible to find the best possible label.
4) Explicit relevance feedback.

2.1 Segmentation And Low Level Feature Extraction

As described in [21] we used texture, color and shapes features:

- Gabor filters to extract texture information, allowing the user to choose the number of scales orientations and sub-images in which the original image is divided.
- colour coherence vector (CCV) which is able to incorporate spatial information and stores the number of coherent pixels for each colour. Coherent pixels belong to an extended region of the same colour
- Fourier descriptors, a contour based technique, to extract shape information that starts with the extraction of the boundary of the object that must be represented. A sequence of ordered points is then selected in order to represent a suitable sample of the boundary following a predefined order. This sequence of the ordered points is parameterized by a signature function. The boundary of the object is represented by applying the chosen signature function to all the boundary pixels and applying the Discrete Fourier Transform (DFT) to the resulting one dimensional function. The scale invariance is obtained by dividing the amplitude of the values of the first half of the Fourier descriptor by the amplitude of the first Fourier Descriptor, as reported in [22]. In order to compare two images we have chosen the Euclidean distance of the feature vector.

2.2 Bayes Point Machine

Using the low level features it is possible to build a classifier. This work presents a supervised classification algorithm that learns how to associate low level features to semantic labels. Using one binary classifier for each label it is possible to obtain a multi-labelling classification that provides the membership probability to each class.

The Bayesan Point Machine (BPM), as proposed by Herbrich in [23] and [24], is used in this study in order to approximate the Bayesian inference for linear classifier in the kernel space.

Given a training set

\[ z = (x, y) = ((x_1, y_1), \ldots, (x_m, y_m)) \in (X \times Y)^m \]

m dimensional, the Bayes classification assigns each new data point \( x \) to the label \( y \) with the minimal expected loss, weighted by its posterior probability:

\[ P_{H|Z} = z(h^*) \]

\[ \text{Bayes}_{z}(x) = \arg \min_{y \in Y} E_{H|Z=x} = z(l(H(x), y)) \]

(2)

where \( l(y, y') \) represent the loss function defined by

\[ l(y, y') = \begin{cases} 1, & y \neq y' \\ 0, & y = y' \end{cases} \]

The solution \( \text{Bayes}_{z} \) in (2) is the one with the minimum loss incurred in each hypothesis \( h \) applied to \( x \) and weighted by its posterior probability. However it is generally impossible to find a unique hypothesis \( h \in H \).
to obtain the solution \( \text{Bayes} \). Herbrich [23] [24], tries to find the closest classifier to the Bayes solution \( \text{Bayes} \) by requiring that the classifier lays inside a fixed hypothesis space \( H \). The approximation \( \text{Bayes} \) is given by:

\[
h_{bp}(z) = \arg \min_{h \in H} E_X \left[ \frac{E_{H|Z^m=x}[l(h(X), H(X))]}{\mathbb{E}_{H|Z^m=x}[l(h(X), H(X))]} \right] = \langle w_{Bayes} X \rangle_F
\]

(3)

where \( w_{Bayes} \) is the weight vector of \( h_{bp}(z) \). The classifier \( h_{bp}(z) \) is called the Bayes point and is the best approximation of the optimal bayes solution \( \text{Bayes} \).

In order to compute (3) it is necessary to know the input training distribution \( P_x \). In [24] it is suggested to find the center of mass in \( V \) with the hypothesis of \( ||w||=1 \) for a two class classification problem, as follows:

\[
w_{cm} = \frac{E_{W|Z^m=z}[W]}{\left\| E_{W|Z^m=z}[W] \right\|}
\]

(4)

Equation 4 can be used to approximate the Bayes point \( h_{bp}(z) \). In [24] in order to compute (4) it is suggested to use a first order Markov chain which can be approximated by uniform sampling weight vectors \( w \in V \) and then by averaging them. In [24] two algorithms are suggested to achieve this sampling: the Billiard-ball algorithm and the perceptron algorithm.

The Billiard-ball algorithm is described in [25], [26]. After initializing the version space with a perceptron or a SVM, the classifier \( w \) behaves like a billiard ball that bounces inside the convex polyhedron \( V \) for a short period of time. This is possible since every training point \( (x_i, y_i) \in Z \) defines an hyperplane

\[
\{w \in W \mid y_i \langle x_i, w \rangle_F = 0 \} \in W
\]

After \( N \) bounces the Bayes Point can be evaluated by means of

\[
w_{cm} = (1/N) \sum_{i=1}^{N} w_i
\]

The previous algorithm has a good generalization performance but a high computational cost. It is possible to apply an approximate uniform sampling of the version space in order to find \( w_{cm} \) on the basis of the fact that each classifier in the version space is already an optimal one.

As described in [27] a Support Vector Machine (SVM) is used to delimitate the version space and the Billiard-ball algorithm to approximate the center of mass.

The aim of the work is to use a set of supervised BPM classifiers, one classifier for each semantic label, to learn the probability that a set of low level features can be associated with a semantic label.

2.3 Relevance feedback

In order to improve the classification result, explicit relevance feedback was used. The user interface allowed the user to check the correctly retrieved image.

For the selected image the value of the probability obtained by the relevant BPM classifier is doubled (up to the unit) while the values of all other labels are halved; for the un-selected image the value of the probability is halved and the first highest probability value among the set of classifier is found and assigned to the image. In this manner it is possible to adjust, by using the user input, the probability of the association between an image and its semantic label.

3. CBIR Description

The CBIR system was developed from scratch in Matlab. The part of the system to extract the low level features was described in [21]. The first part of the new system was developed to build the training data and associating one semantic label to each image of the training set.

Figure 1 shows the interface. It is possible to associate an image to a label by means of the File panel, to create new labels by means of the Label panel, and to manage the image database by means of the Database panel.

The second interface showed in Figure 2 allows one to perform the classification step. The interface allows one to browse the database and select an image, to choose the low level features that will be used for the classification phase, to perform a query by example or a query by label, to show the query results in page of 15 images each with the possibility for the user to select the relevant images thus allowing for the relevance feedback and to show the first five labels assigned to each image ordered by decreasing probability.

After selecting the low level features that will be used in the classification phase the start button is enabled and the classification phase can begin.

The algorithm start by constructing a matrix containing the low level features in accordance with the
user selection. If, for example, the user choose to use only texture, the input matrix will be composed of 1,184 columns. The low level features are normalized. In the training phase a binary output vector is created for each classifier. The dataset is divided in a training and testing phase. The dataset are used by the BPM classifier. Using the Bayes Point Machine toolbox in Matlab [28] a classifier is built, trained, tested and used.

Figure 1. Tool interface for label management.

After the training phase is performed the user can perform a query by example or a query by label. Figure 3 shows the results obtained for a query for all the image belonging to the Latero Lateral cephalometric x-rays.

In this case the tool was able to retrieve all relevant images in the first 15 results among the 45 x-rays present in the dataset. The classifiers used only 2 x-rays in the training phase.

Figure 2. Tool interface for the classification.

4. CBIR System Evaluation

The testing of the system was performed with 225 images 135 in medical domains and 90 in other domains. For the importance of Latero Lateral (LL) cephalometric analysis the reader can reference [29-30] for automatic landmarking process, and [31] for the landmarking reliability. The reliability of Postero Anterior (PA) cephalometric analysis is assessed in [32][33].

All the images were 24 bitmap with a fixed dimension of 256 x 256 pixels. When necessary the images were resized to this dimension preserving their ratio.

Table 1 shows the percentage of correctly retrieved images among the first $X + 10\%$, where $X$ is the total number of images present in the database with the corresponding label. Hence, if in the database there are $X$ images with the given label the table reports the number of correct images retrieved among the first $X + 10\%$ images.

The results in the third column were obtained when gray scale, Gabor and shape were used as low level features; the results in the fourth column were obtained with shape and Gabor as low level features and the value in the last column were obtained with gray scale and shape as low level features.

From the table it is possible to highlight the good results for all the x-rays hence for all the images related to the medical domain in this study. In particular the best performance for hand x-rays, 85% of correct results over the first 50 images, is obtained by using either gray scale, Gabor and shape or gray scale and shape thus resulting in a high influence of the shape features. The best performance for Latero Lateral x-rays, 84 % of correct results over the first 50 images, is obtained by using shape and Gabor as low level features. The best performance for the Postero Anterior x-rays, 92% of
correct results over the first 50 images, is obtained by using shape and Gabor as low level features.

Overall for the medical image the shape and Gabor low level feature have the best discriminating factor, followed by colour.

During this work other combinations of low level features have been experimented: grey scale, shape and colour; grey scale and shape. The analysis of the results shows a quite good performance, but not as good as the reported case, by the shape feature.

The results are in line with other medical images retrieval system based on labels such as the work presented in [34].

<table>
<thead>
<tr>
<th>Features</th>
<th>% among X + 10%</th>
<th>% among X + 10%</th>
<th>% among X + 10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Tree</td>
<td>13</td>
<td>47</td>
<td>13</td>
</tr>
<tr>
<td>2 Flower</td>
<td>43</td>
<td>33</td>
<td>43</td>
</tr>
<tr>
<td>3 Snow</td>
<td>8</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>4 Hand</td>
<td>85</td>
<td>83</td>
<td>85</td>
</tr>
<tr>
<td>5 LL RX</td>
<td>52</td>
<td>84</td>
<td>52</td>
</tr>
<tr>
<td>6 PA RX</td>
<td>85</td>
<td>92</td>
<td>85</td>
</tr>
<tr>
<td>7 Pyramid</td>
<td>45</td>
<td>50</td>
<td>45</td>
</tr>
<tr>
<td>8 Sky</td>
<td>25</td>
<td>13</td>
<td>25</td>
</tr>
<tr>
<td>9 Building</td>
<td>31</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>10 Bird</td>
<td>18</td>
<td>9</td>
<td>18</td>
</tr>
</tbody>
</table>

For all other images the classification was obtained by using by Gabor, color and shape features. The not so good performance for the non-medical images is due to the low number of image in the database and hence the low number of training examples.

Among the non-medical images someone such as pyramid are better retrieved due to the characteristic shape. The best performance in this case is obtained with the shape and Gabor low level features.

This work shows how a careful choice of tools and algorithms and the application of software engineering techniques it is possible to obtain a fully working prototype of a complex software, such as a CBIR system. The prototype can be used to experiment with the implemented algorithms and techniques and allowing to extend the system with other algorithms.

5. Conclusion

In this work we have presented an in-house developed CBIR system designed and developed from scratch using Matlab. The tool is able to perform both query by example and query by label hence to retrieve semantically relevant images. The semantic image retrieval is based on a supervised BPM classifiers. A first evaluation of the CBMS in the medical domain is presented.

As further study we plan to deeply evaluate the CBMS tool with a greater number of x-rays and a greater number of parameter to evaluate its performance.

The CBMS can also be integrated with information automatically extracted and classified from public database such as PUBMED as described in [35].

References

[10] A. Rosset, Casimage Project: Integration of a Multimedia Teaching and Reference Database in a PACS Environment, Thèse présenté à la Faculté de Médecine de
l'Université de Genève pour obtenir le grade de Docteur en Médecine.


[21] F. Maiorana, & C. Pino, Design implementation and evaluation of two medical content based image retrieval systems, Proceedings of the Ninth IASTED International Conference on Biomedical Engineering (BioMed 2012), Innsbruck, Austria, 2012,


[28] Bayes Point Machine Toolbox
http://research.microsoft.com/en-
us/um/people/minka/papers/ep/bpm/


