Image Recognition for Digital Libraries [*]

Bertrand Le Saux
ISTI - CNR
Via G. Moruzzi, 1 - 56124 - Pisa - Italy
bertrand.lesaux@isti.cnr.it

Giuseppe Amato
ISTI - CNR
Via G. Moruzzi, 1 - 56124 - Pisa - Italy
giuseppe.amato@isti.cnr.it

ABSTRACT

The interpretation of natural scenes, generally so obvious and effortless for humans, still remains a challenge in computer vision. To allow the search of image-based documents in digital libraries, we propose to design classifiers able to annotate images with keywords. First, we propose an image representation appropriate for scene description. Images are segmented into regions, and then indexed according to the presence of given region types. Second, we propound a classification scheme designed to separate images in the descriptor space. This is achieved by combining feature selection and kernel-method-based classification.

General Terms

Algorithms

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries

Keywords

Scene analysis, image segmentation, clustering, feature selection, image classification, kernel-method

1. INTRODUCTION

Meta-data describe the content of documents managed by digital libraries and are used for searching the documents themselves. Simple flat meta-data such as the Dublin Core [1] cannot satisfy the requirements of advanced multimedia digital libraries, for which an elaborate description of the visual content is needed. Thus, we are observing the development of increasingly complex and structured meta-data models [3, 10, 24] which often include nested structures, hierarchies, multiple views and semi-structured information. On one hand, the adoption of complex meta-data models allows cataloguers to generate very precise descriptions of the documents, but on the other hand, the cost of generating such descriptions is much higher than with simple meta-data models.

The use of techniques for automatic generation of meta-data, integrated with manual generation and validation of meta-data, seems to be a promising approach that allows, at the same time, the use of complex meta-data and acceptable costs for producing them [3].

This paper proposes a novel technique for automatic generation of meta-data for image-based documents. It annotates images with predefined semantic concepts by combining classification methods and techniques of image processing and visual feature extraction.

Images may be indexed in various ways to stress some specific visual information. Those different approaches are extensively studied in the field of content-based image retrieval [15, 22]. Previously, image classification has been performed by using directly support-vector machines on image histograms [11] or hidden Markov models on multi-resolution features [23]. Since the human description of an image content is often specific to an image part or an object, approaches using blobs to focus on local characteristics have been propounded to capture higher-level information and then used for classification [7, 16]. The focus on key-regions was also shown to be promising to perform an image search: in [25], image blocks were associated in groupings of similar regions across images to learn various concepts and thus retrieve images by relevance feedback. To allow semantic queries based on icons - such as people, buildings, trees - a visual-concept detection system based on feature selection was proposed in [21].

We believe that a segmentation of images into regions can be used efficiently to provide more semantic information than the usual global image features. Therefore, we first propound to identify which region types are present in an image. Hence, the images that contain the same region types are likely to be associated with a particular semantic concept. Then we propose to define a scene classifier that can test the co-presence of these region types. We also adopt a feature selection step to avoid over-fitting on meaningless region types.

This paper is organized as follows. Section 2 describes how the scene information is represented by the means of *presence vectors*. We explain how we design classifiers able to

[*] This research was partially funded by the ECD (Enhanced Content Delivery) project and the Delos Network of Excellence.

MIR'04, October 15–16, 2004, New York, New York, USA.

associate image and concept in section 3. Section 4 explores the mechanism of the classification process and evaluates the proposed method. The use of the technique in a system for managing digital libraries is discussed in section 5.

2. FEATURE EXTRACTION

The analysis of images is based on a segmentation of each image into visually-significant regions. The image regions are compared to a *region lexicon* to obtain a *presence vector* which describes the region types present in the image.

2.1 Image segmentation

This segmentation is processed using the mean-shift algorithm [13]. It is a simple non-parametric technique for maximization of the probability density, by performing basically a density gradient ascent. To perform color segmentation, the mean-shift procedure is applied at various start locations in the color space, then the obtained high-density colors are mapped to the image plane to keep only those belonging to large-enough regions.

The segmentation is tuned to produce between three and ten regions per image, depending on the complexity of the scene. These regions are not connected but correspond to areas with the same dominant color (cf. figure 1).

2.2 Region lexicon

The region lexicon consists in a discrete range of regions that occur in an image dataset. Such a dataset is built by gathering various generic images. Once they are segmented, these images are assumed to provide a good overview of the possible image regions that occur in the real world.

We clusterize this data-set of image regions using techniques previously used to find clusters of visually-similar images in image databases [20] and based on fuzzy clustering methods [8]. Since the regions are more homogeneous than whole images, mere descriptors like the mean color can be used instead of color histograms to deal efficiently with the problems due to the complexity of clustering algorithms. The resulting clusters contain visually-similar image regions and define a region type. The region lexicon is formed by these region types.

2.3 Presence vectors

Given a discrete set of region types, obtained as described above, every image can be described by a presence vector: each component corresponds to a region type and its value can be true or false depending on the fact that the region type is present or not in the image.

The decision on the presence of a region type is taken by measuring the similarity - in the visual-descriptor space - between the cluster centroid and the regions of the image. For instance, the image of figure 1 contain skin-colored or greenery regions.

3. IMAGE CLASSIFICATION

We aim to build binary classifiers able to associate a concept to a given category of images. Typically, these classifiers can be used to annotate images with keywords. The keyword is basically a mnemonic representation of a concept such as people, countryside, etc.

The classification scheme is a two-step process. First a feature selection allows to determine which typical regions are important for a given concept. Second a kernel classifier is used to learn a decision rule from the selected region types.

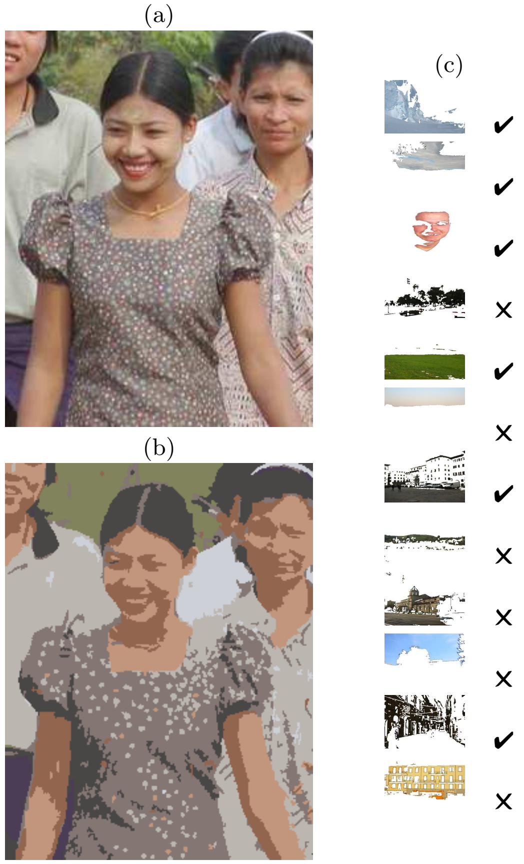

Figure 1: Feature extraction: the original image (a) is first segmented (b) then the corresponding boolean presence vector (c) is extracted by comparing the image regions to the typical ones obtained from the clustering of a training set.

Let $\mathcal{I}$ denote the set of images, and X a random variable on $\mathcal{I}$ standing for the distribution of images. We denote Y a boolean random variable for the class to predict, i.e. the scene keyword to associate with the image.

We consider a set of features $F = \{f_1, \ldots, f_N\}$ which are mappings from $\mathcal{I} \rightarrow \{0, 1\}$. In the experiments those features are indicators of the presence - or absence - of a given region type in the image. We denote $F_1 = f_1(X), \ldots, F_p = f_p(X)$ the boolean random variables associated with those features.

3.1 Feature selection

In this application, the selection of important features is a filtering phase to understand which region types are meaningful to recognize a concept. A review of various feature selection (FS) techniques is provided in [19]. The most standard ways to select features consist in ranking them according to their individual predictive power, that may be estimated by mutual information [6].

Information theory [18] provides tools to assess the available features. The entropy measures the average number of bits required to encode the value of a random variable. For instance, the entropy of the class Y is $H(Y) = -\sum_y P(Y =$

Algorithm 1 regularized kernel-adatron with bias

```
1. Initialize $\alpha_i = 0$ and $b = 0$
2. For all labeled point $(x_i, y_i)$ calculate:
   $$z_i = \sum_{j=1}^{p} \alpha_j y_j K(x_i, x_j) - b$$
3. Calculate $\delta\alpha_i = (1 - y_i z_i)/K(x_i, x_i)$ the proposed change to the multipliers $\alpha_i$
4. Update $\alpha_i$
   (a) If $\alpha_i + \delta\alpha_i \leq 0$ then $\alpha_i = 0$ and $b \leftarrow b - y_i\alpha_i$
   (b) If $0 < \alpha_i + \delta\alpha_i \leq C$ then $\alpha_i \leftarrow \alpha_i + \delta\alpha_i$ and $b \leftarrow b + y_i\delta\alpha_i$
   (c) If $\alpha_i + \delta\alpha_i > C$ then $\alpha_i = C$ and $b \leftarrow b + y_i(C - \alpha_i)$
5. Stop if stability is reached, otherwise return to step 2
```

$y)\log(P(Y = y))$. The conditional entropy $H(Y|F_j) = H(Y, F_j) - H(F_j)$ quantifies the number of bits required to describe Y when the feature F_j is already known. The mutual information of the class and the feature quantifies how much information is shared between them and is defined by:

$$\begin{aligned} I(Y, F_j) &= H(Y) - H(Y|F_j) \qquad (1) \\ &= H(Y) + H(F_j) - H(Y, F_j) \end{aligned}$$

The features f_j are ranked according to the information $I(Y, F_j)$ they convey about the class to predict. Those with the largest mutual informations are chosen.

3.2 Kernel-adatron classifiers

The adatron was first introduced in [5] as a perceptron-like procedure to classify data and a kernel-based version was then proposed [17]. Basically, it performs a gradient ascent to solve the margin-maximization problem between the two classes of the training set and thus is a simple implementation of a support-vector machine [26].

The training dataset is denoted $T = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ independently generated from (X, Y). Each x_i corresponds to a vector of realizations of the selected features F_j - it is a presence vector from which the meaningless components have been cut off - and y_i is true or false according to the fact that the training image x_i is a positive or negative sample of the concept to learn. The algorithm tries to estimate the parameters α_i and b of the decision function based on the chosen kernel $K(.,.)$:

$$f(x) = sign(\sum_{i=1}^{n} y_i \alpha_i K(x, x_i) + b) \qquad (2)$$

We use a regularized version of the method: the α_i are constrained under a constant value C to prevent over-fitting to a particular vector. It is summarized in algorithm 1.

4. EXPERIMENTS

4.1 Data-set

The first data-set is composed of 5 classes of images. Four classes contain instances of a particular scene: *snowy, countryside, streets* and *people*. The fifth one consists of various generic images aimed to catch a glimpse of the possible real scenes and thus used as negative samples for the classifiers.

The second data-set comes from a news agency. In this context the concepts to learn are news topics: the database was divided in categories *politics* and *sport* and another one of negative samples.

In the experiments, training categories of 30 instances are extracted randomly from the data-sets and error rates are averaged on 50 runs.

4.2 Error rates

First we aim to test the validity of the image representation by presence vectors. A linear classifier tests only the co-presence of the region types to attribute a given label. The results (cf. table 1) show that even with this mere classifier the description scheme is efficient enough to separate different categories.

Then, the use of a polynomial-kernel adatron enables to obtain smaller error rates, but over-fitting on the training data still impedes the correct classification of the more complex scene as *people* or *streets*, for which the variance of the possible regions is greater.

Finally, by selecting the most informative region types for each category (column "FS+polynomial adatron"), we can eliminate those which are not relevant, for instance some background details. In this manner, performances on complex scenes are as good as those on simple ones.

4.3 Meaningful regions

Since we know the region types that are meaningful for a given keyword, it is possible to retrieve which parts of the image are taken into account to recognize a given keyword. Figure 2 shows images from which the meaningless regions have been cut off. Those images come from the news-agency dataset and are processed by classifiers trained on the two main possible topics: *sport* and *politics*.

The selected region types are consistent with what could be intuitively expected. For the tennis-player photo, the selected regions for the *sport* keyword correspond to the green walls of the playground and to the bright jerseys whereas those considered as interesting for the *politics* keyword are details of the background the color of which could have been significant. For the politician image, the results are opposite. The feature selector for *sport* selects the regions from the gold background. Though they cover a large part of the image, they are not considered as relevant enough to give the label. On the contrary, the face and dark-suit regions are meaningful for *politics* and moreover relevant to the classifier.

Actually, the region types selected by mutual information maximization can be either regions that are representative of the training category or on the contrary regions that occur often in the negative samples. Figure 3 shows which regions among the selected ones are positive clues for the classifier. For instance, sky regions have not a highest probability to appear in the training images for *countryside* than on the whole dataset, to the contrary of greenery regions for which $P(F_i \mid Y) > P(F_i)$. Using Bayes' rule, these last ones are typical of a particular scene - i.e. $P(Y \mid F_i) > P(Y)$ - and correspond to the model built by the classifier to represent it. Positive clues for *countryside* are green and dark-brown regions, and for *snowy* they are white and blue-sky regions.

Table 1: Error rates for various keywords: comparison of various classification schemes applied to presence vectors

keyword	linear adatron		polynomial adatron		FS + polynomial adatron	
	training error	test error	training error	test error	training error	test error
snowy	0.0 %	9.2 %	2.3 %	8.9 %	2.4 %	8.5 %
countryside	0.0 %	12.6 %	0.0 %	9.1 %	8.0 %	8.4 %
people	3.6 %	16.4 %	0.5 %	14.1 %	3.6 %	7.5 %
streets	0.1 %	14.0 %	0.0 %	12.1 %	2.5 %	6.2 %

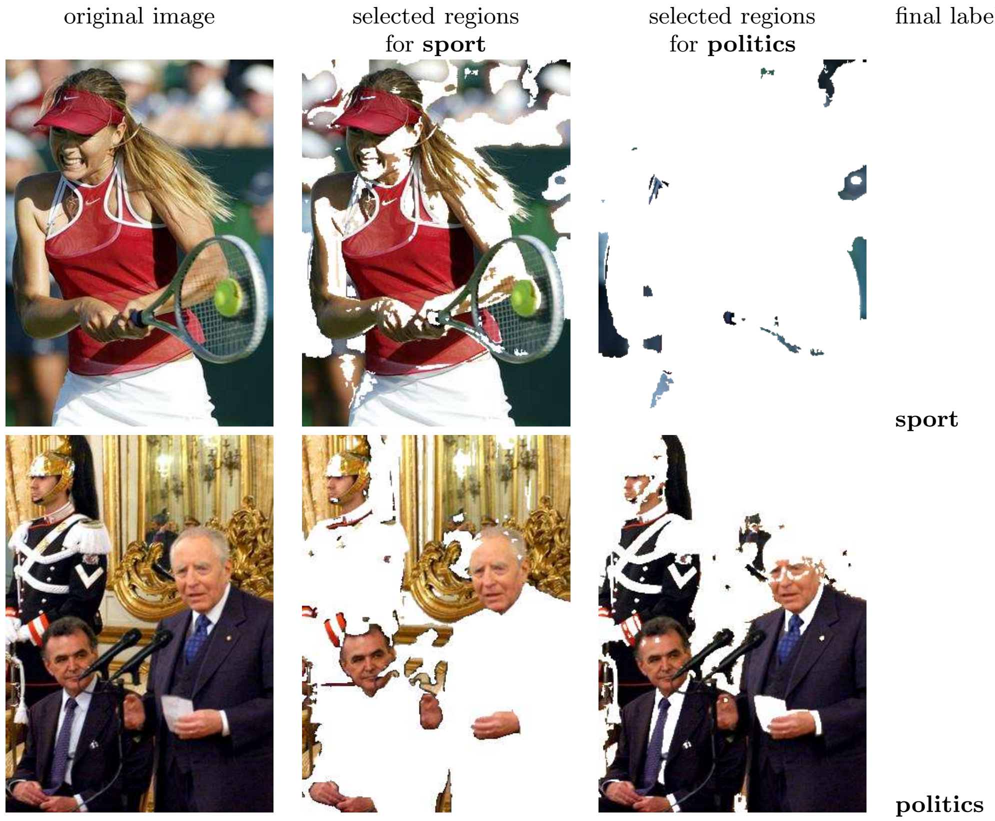
original image
selected regions for **sport**
selected regions for **politics**
final label
sport
politics

Figure 2: Meaningful regions used to recognize news topics: the original images are presented in the first column, the image regions which were considered as meaningful to classify the image according to keywords *sport* and *politics* are shown in the second and third ones respectively, and the last one shows the keyword the image was finally labeled with.

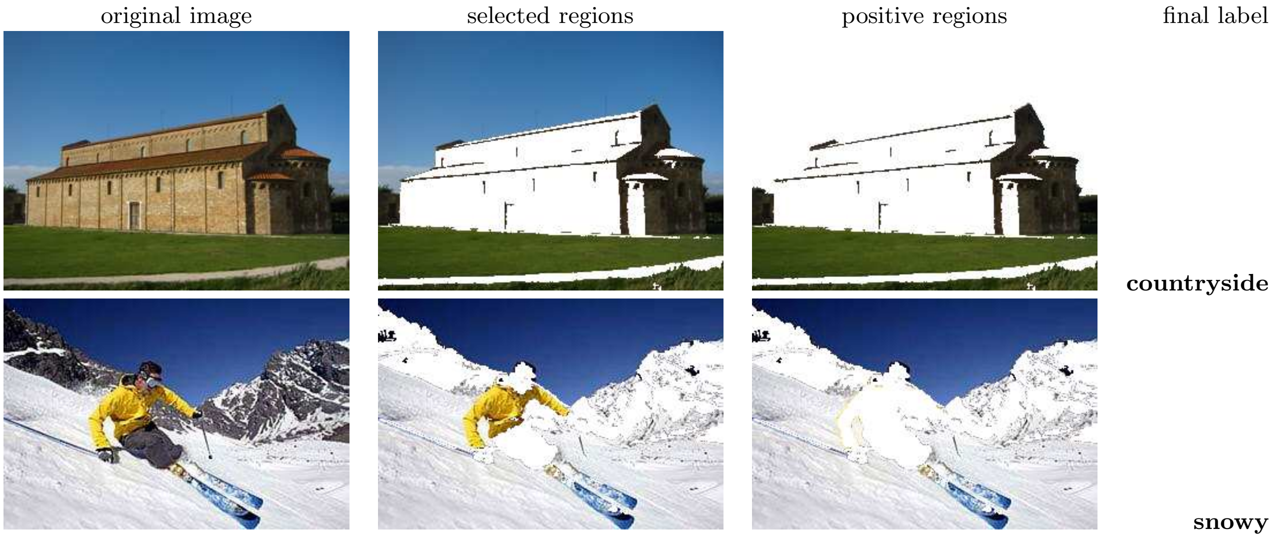

Figure 3: Meaningful regions used to recognize natural scenes. The meaningful regions correspond to the region type whose mutual information with the label to predict is high. Among them, the positive regions are the ones the type of which is a positive clue to label the image, i.e. $P(Y \mid F_i) > P(Y)$.

4.4 Comparison

Our approach is compared with a SVM applied to image histograms [11]. The error rates for both methods are shown in table 2. The histogram-based approach works well for the simple scenes that likely have high-density peaks on some colors. However, the performances are less effective for the more complex types of scene: various backgrounds make the generalization harder.

On the contrary, our classifiers used on presence-vectors obtain roughly the same error rates for all kinds of scene. On the complex ones, both the segmentation into regions and the feature selection allow to catch the details that permit to differentiate these images from others without over-fitting.

5. DIGITAL LIBRARY INTEGRATION

The image classification tool has been integrated in the MILOS system [4]. MILOS is a Multimedia Content Management system specialized to support document intensive applications as for instance Digital Libraries. In a few words, MILOS plays with document intensive application the same role played with data intensive applications by databases. MILOS provides the developer of document intensive applications with functionalities for efficient and effective management of multimedia documents and their meta-data.

MILOS is composed of three main components as depicted in Figure 4: the Meta-data Storage and Retrieval (MSR) component, the Multi Media Server (MMS) component, and the Repository Meta-data Integrator (RMI) component. All these components are implemented as Web Services and interact by using SOAP. The MSR manages the meta-data used by applications. It relies on our technology for native XML databases. The MMS manages the multimedia documents used by applications. The RMI implements the service logic of the repository providing developers of applications with a uniform and integrated way of accessing MMS and MRS. In addition, it supports the mapping of different meta-data schemas.

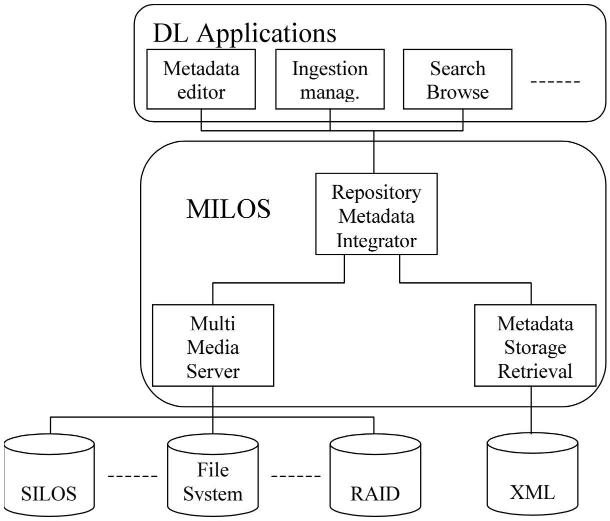

Figure 4: General Architecture of MILOS

An important functionality required by document intensive applications, and offered by MILOS trough the MSR, is the search functionality. Searches are typically performed on the meta-data, which describe the content, and the bibliographic information of the managed documents.

MILOS supports searches by relying on an enhanced native XML database system which offers special functionalities needed by document intensive applications. The use of a native XML database is especially justified by the well known and accepted advantages of representing meta-data as XML documents. XML-represented meta-data might have arbitrary complex structures, which allow to deal with complex meta-data schemes, and might be easily exported and imported.

The new generation of XML encoded meta-data standards, such as MPEG-7, include in their description fea-

Table 2: Error rates for various keywords: SVM on histograms vs. feature selection and polynomial adatron on presence vectors

keyword	SVM on histograms		FS + polynomial adatron	
	training error	test error	training error	test error
snowy	0.0 %	5.0 %	2.4 %	8.5 %
country	0.0 %	9.5 %	8.0 %	8.4 %
people	0.0 %	13.2 %	3.6 %	7.5 %
streets	0.1 %	15.2 %	2.5 %	6.2 %

tures automatically extracted from visual documents, such as color histograms, textures, shapes, etc. Thus our XML database supports high performance search and retrieval on heavily structured XML documents, relying on specific index structures [2, 27], as well as full text search, automatic classification [12], and feature similarity search [9].

The MILOS XML database allows the system administrator to associate specific XML element names with special indexes. Therefore, for instances, the tag name `<abstract>` can be associated with a full text index and the MPEG-7 `<VisualDescriptor>` tag can be associated with a similarity search index structure.

From this perspective, the image classification tool described in this paper can be seen as another special index to be used by the XML database. To this aim we have defined an XML encoding of the presence vectors described in Section 2. The MILOS administrator can associate tag names (provided that they contain the XML encoding of the presence vectors) with the *image classification index*. When the XML database is asked to insert a new XML document, if the specified tag is encountered, the tag's content is passed to the image classification index, which classifies the received presence vector using the available classifiers. Various classifiers can be generated by using the *learning tool* described in Section 3. A possible work-flow followed by an application that implements an image digital library can be as follows: 1) the application receive the image to be inserted in the digital library; 2) the application gets some meta-data directly from the user; 3) the application automatically generates additional XML encoded meta-data, by using various image processing tools (provided by MILOS), including the presence vector generator 4) the application merges manual and automatically generated meta-data and encodes everything in XML 5) the application inserts the XML meta-data into the XML database.

To deal easily and transparently with these advanced search and indexing functionalities, we have extended the syntax of the basic XQuery language [14] with new operators that also deal with classification categories.

Consider for example the following query :

```
for $a in /images
where
    category($a/pres-vec)= 'sport' and
    $a/provider = 'Reuters' and
    $a/production-date = '2003/11/12' and
    $a/VisualDescriptor~
    MPEG-7-extract('http://www.there.com/image7.jpg')
return $a
```

It searches for the images classified as 'sport', provided by 'Reuters', produced on '2003/11/12', and visually similar to a picture available on the web. Similarly, figure 5 provides the first results of the MILOS search engine for a simpler query based on the concept *sport*.

6. CONCLUSION

We have presented in this article a new approach to scene recognition that intends to identify the image-region types that compose one scene. It has been implemented in a digital library system to perform automatic image annotation and thus to allow textual queries of image-based documents.

The strong points of the proposed approach are the following:

- the image representation is particularly appropriate for scene description, just as the classification scheme, since the region-type selection step lets find the regions that are visually relevant to distinguish a given scene;
- it provides a good tradeoff between the classifier performance and the prevention of over-fitting;
- it is robust, as it does not require a fine tuning of a complex algorithm but at the contrary uses a succession of simple procedures.

This method is valid for complex scenes, as soon as there is one region type distinctive enough. Nevertheless, such an image description is certainly too rough to distinguish between look-alike scenes, especially with the sole color. Forthcoming works will investigate further how we can refine the model of the scene, for example by taking into account the spatial arrangement of the regions.

7. REFERENCES

[1] Dublin Core Metadata Initiative. http://dublincore.org.

[2] G. Amato, F. Debole, F. Rabitti, and P. Zezula. YAPI: Yet another path index for XML searching. In *European Conference on Digital Libraries*, Trondheim, Norway, August 2003.

[3] G. Amato, C. Gennaro, and P. Savino. Indexing and retrieving documentary films: managing metadata in the ECHO system. In *4th Intl. Workshop on Multimedia Information Retrieval, in conjunction with ACM Multimedia*, Juan-les-Pins, France, December 2002.

[4] G. Amato, C. Gennaro, P. Savino, and F. Rabitti. Milos: a multimedia content management system for digital library applications. In *European Conference on Digital Libraries*, Bath, U.K., september 2004.

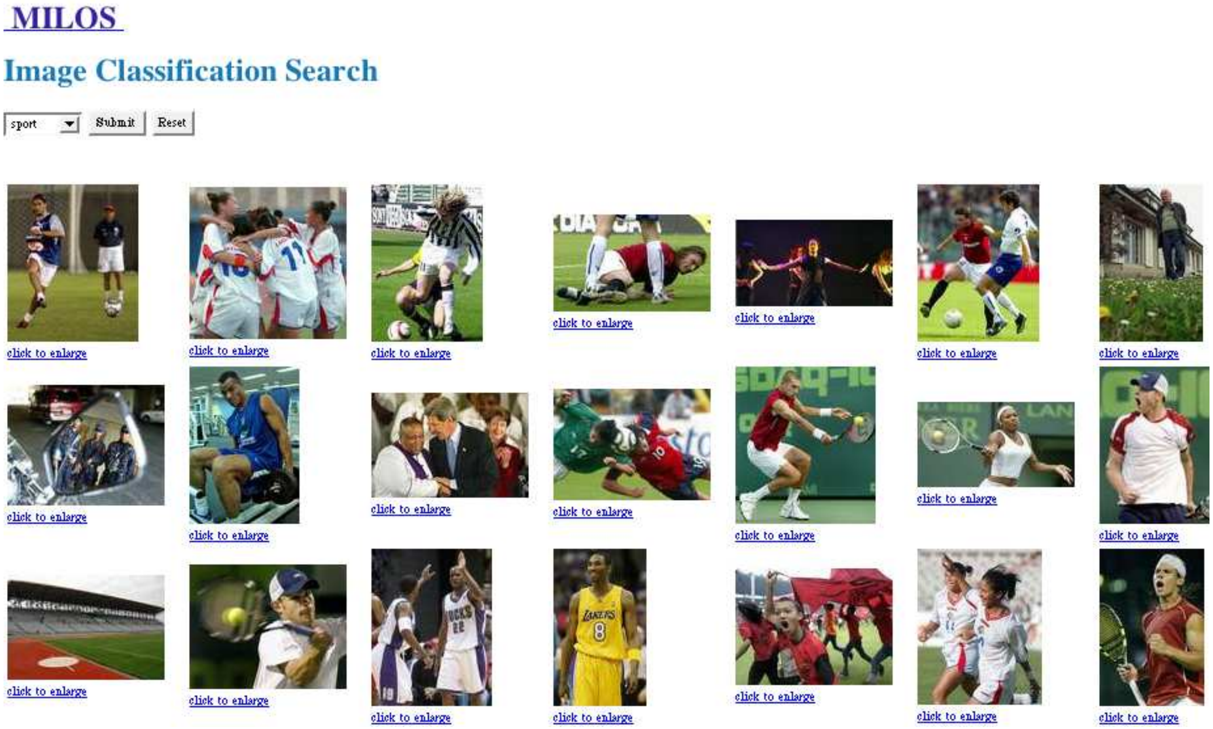

Figure 5: MILOS interface for query on images automatically annotated with concepts.

[5] J. Anlauf and M. Biehl. The adatron: an adaptive perceptron algorithm. *Neurophysics Letters*, 10:687–692, 1989.

[6] R. Battiti. Using mutual information for selecting features in supervised neural network learning. *Neural Networks*, 5(4):537–550, 1994.

[7] S. Belongie, C. Carson, H. Greenspan, and J. Malik. Recognition of images in large databases using a learning framework. Technical report, University of California at Berkeley, 1997.

[8] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New-York, N.Y., 1981.

[9] C. Böhm, S. Berchtold, and D. Keim. Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. *ACM Computing Surveys*, 33(3):322–373, September 2001.

[10] D. Castelli and P. Pagano. Opendlib: A digital library service system. In M. Agosti and C. Thanos, editors, *European Conference on Digital Libraries*, Rome, Italy, September 2002.

[11] O. Chapelle, P. Haffner, and V. Vapnik. Svms for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10:1055–1065, 1999.

[12] N. Christianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[13] D. Comaniciu and P. Meer. Robust analysis of feature spaces: Color image segmentation. In *Conference on Computer Vision and Pattern Recognition*, pages 750–755, San Juan, Porto Rico, June 1997.

[14] W. W. W. Consortium. XQuery 1.0: An XML query language. W3C Working Draft, November 2002. http://www.w3.org/TR/xquery.

[15] A. Del Bimbo. *Visual information retrieval*. Morgan Kaufmann, San Francisco, CA, 1999.

[16] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision*, volume 4, pages 97–112, Copenhagen, Denmark, May 2002.

[17] T.-T. Friess, N. Christianini, and C. Campbell. The kernel-adatron algorithm: a fast and simple learning procedure for support vector machines. In *International Conference on Machine Learning*, Madison, Wisconsin, July 1998.

[18] R. M. Gray. *Entropy and Information Theory*. Springer-Verlag, New York, N.Y., 1990.

[19] I. Guyon and A. Elisseff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[20] B. Le Saux and N. Boujemaa. Unsupervised robust clustering for image database categorization. In *International Conference on Pattern Recognition*, Quebec, Canada, August 2002.

[21] M. Lew. Next generation web searches for visual content. *IEEE Computer*, pages 46–53, 2000.

[22] M. Lew. *Principles of Visual Information Retrieval*. Springer-Verlag, London, U.K., 2001.

[23] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistic modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1075–1088, 2003.

[24] B. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, New-York, N.Y., 2002.

[25] T. Minka and R. Picard. Interactive learning using a society of models. *Pattern Recognition*, 30(4):565–581, 1997.

[26] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New-York, N.Y., 1995.

[27] P. Zezula, G. Amato, F. Debole, and F. Rabitti. Tree signatures for xml querying and navigation. In *Database and XML Technologies, First International XML Database Symposium*, pages 149–163, Berlin, Germany, September 2003.