Use of Weighted Visual Terms and Machine Learning Techniques for Image Content Recognition relying on MPEG-7 Visual Descriptors

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

Pasquale Savino
ISTI-CNR
Via G. Moruzzi, 1
56124, Pisa, IT
pasquale.savino@isti.cnr.it

Vanessa Magionami
ISTI-CNR
Via G. Moruzzi, 1
56124, Pisa, IT
vanessa.magionami@isti.cnr.it

ABSTRACT
We propose a technique for automatic recognition of content in images. Our technique uses machine learning methods to build classifiers which are able to decide about the presence of semantic concepts in images. Our classifiers exploit a representation of images in terms of vectors of visual terms. A visual term represents a set of visually similar regions that can be found in images. Various types of visual terms are used at the same time to take into account various similarity criteria and region representations that are available to compare regions. Specifically, we compare regions using the 5 MPEG-7 visual descriptors. An image is indexed by first using a segmentation algorithm to extract its regions, and then the image is associated with the visual terms that are more similar to the extracted regions. The proposed technique offers very good performance as demonstrated by the experiments that we performed.

Categories and Subject Descriptors:
H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval;

General Terms
Measurement, Performance, Experimentation

Keywords
Image content classification, machine learning

1. INTRODUCTION
In this paper we propose a new technique for automatic image content annotation. Our technique is based on the use of automatic classifiers to determine the content of predefined concepts, as for instance car, person, landscape, flower, skyline, countryside, sport, etc, in an image. We use Support Vector Machines (SVM) [5] to build classifiers, applied to a special representation of images.

MS'08, October 31, 2008, Vancouver, British Columbia, Canada.

Humans decide about the occurrence of a concept in an image on the base of the co-occurrence of various combinations of visual regions. Accordingly, we represent the visual content of images in terms of visual regions. More precisely, we do not associate images with their actual regions; rather the original regions are represented by prototypes of regions taken from a controlled set of typical regions. We call the controlled set of typical regions, used to describe the content of images, the *visual lexicon* and each typical region in the visual lexicon a *visual term*. We use visual similarity between original regions and regions of the visual lexicon to determine which regions have to be associated with an image. This technique to represent (to index) the content of an image is similar to those presented in [6][9]. The main innovation of our approach, as discussed later, is that we define several sets of visual terms, each one containing visual terms built according to different MPEG-7 visual features. For instance there will be a set of visual terms that is based on colour histograms, one based on textures, one based on shapes, etc.

Visual terms associated with an image are also associated with a weight that specifies the relevance of the visual term in the image. Later, we will see that this weight is determined according to statistical properties of the data sets, concepts to be recognized, and to the similarity between the actual regions contained in the image and the visual terms. This also represents a difference with respect to the indexing approaches presented in [6][9], where association between visual terms and images is binary and not weighted.

Note that the occurrence of a single visual term in an image does not imply any semantic meaning in an image. It just means that the image contains a region very similar to it. It is the combination of occurrences of several visual terms that suggests the presence of a certain semantic content in the image.

In order to build a classifier, learning algorithms for SVM require a training set composed of positive and negative examples of the class to be recognized. Provided that data to be classified are represented by vectors, an SVM classifier is basically able to distinguish the portions of the space containing positive examples (occurrence of the concept) from those containing negative examples. In our context, the vectors representing images are sequences of weighted visual terms (unassigned visual terms are implicitly assigned weight 0). Positive examples represent images containing a certain concept, while negative examples represent images where the concept does not occur.

The paper is organized as follows. Section 2, describe the techniques for generating the visual lexicon, that is the set of

visual terms used to represent visual content of images. Section 3 presents the technique used for associating visual terms with images. Section 4 deals with the issue of setting up classifiers for recognizing image content. Section 5 discusses the experiments that we have performed to assess the performance of our approach. Section 6 concludes.

2. VISUAL LEXICON GENERATION

Our approach indexes images by using visual terms, consisting of typical visual regions that can occur in images. The occurrence of a particular combination of regions is used by a classifier to decide about the presence of a concept in an image. Clearly, creating a visual vocabulary that contains all possible regions is not a good idea. However, as also discussed in [9][1][6], very similar regions play the same role, from a visual point of view, in the process of contributing to form a concept in the image. Assuming that visual similarity is an indication of equivalence of the role of visual regions, we represent a group of very similar visual regions by using a single region representative, which is chosen from the visual lexicon.

Visual similarity can be judged according to several aspects. In fact, in the literature, several visual features and correspondingly several similarity measures have been defined to characterize visual similarity. For instance, MPEG-7 [11] defines 5 visual descriptors, corresponding to global colours, scalable colours, hedge histograms, homogeneous textures, and region shape.

We propose to use a multi-feature visual lexicon, obtained as a composition of several mono-feature visual lexicons. Visual terms in a mono-feature visual lexicon are used to represent regions visually similar according to one specific feature. A multi-feature visual lexicon contains terms, belonging to different mono-feature lexicons, which are able to represent types of regions according to different visual similarity criteria.

In order to build the visual lexicon we start with a training set of images, which are representative of the dataset of images we want to deal with. We apply a region segmentation algorithm to this set of images and we obtain a set of regions. Provided that the image training set has been properly chosen and the region segmentation algorithm identifies good regions, we will end-up with a representative set of regions of the dataset. At this point we have to reduce the size of the region set in order to have the visual lexicon. In this work we have tested two different strategies.

The first very simple strategy consists in defining the size of the visual lexicon and to randomly choose the visual terms out of the region set obtained after segmentation.

For the second strategy we tried to apply the k-medoids [8] clustering algorithm to group together visually similar regions. Cluster representatives are the visual terms. An application of the clustering algorithm using a specific visual feature and similarity function generates a mono-feature visual lexicon. Several applications of the clustering algorithm using different visual features generate the multi-feature visual lexicon.

3. IMAGE INDEXING

A visual term is associated with an image when the image contains a region very similar to the visual term. This is obtained through the following three steps: 1) image segmentation, 2) visual terms selection, 3) visual terms weighting.

Given an image to be indexed, a region segmentation algorithm is used to identify the set of visual regions that appear in it. The selection of the relevant visual terms is obtained by choosing for each extracted region a visual term to be used to represent the region. In order to do that, each region is associated with the most similar visual term in each mono-feature lexicon. Therefore, if the visual lexicon is composed of n mono-feature lexicons, each region extracted from the image will be represented by n visual terms.

After we associate an image with a set of visual terms, we have to decide the relevance (weight) of each associated visual term in the image. The importance of a visual term for an image can be obtained with a measure very similar to what in text retrieval is called the Term Frequency (*TF*) of a term in a document. In text retrieval the term frequency TF^D_t of a term t in a document D is directly proportional to the number of occurrences of t in D. In our context, intuitively, the importance of a visual term t in an image I should be directly proportional to the similarity between a region r of the image and the visual term and to the size of the region in the image. In addition, it should be directly proportional to the number of regions, in the image, represented by t. This can be expressed as

$$TF_t^I = \sum_{r \in Regions(I,t)} sim(r,t) * cover(r,I)$$

where *Regions(I, t)* is the set of regions of I that are represented by t, *sim(r, t)* gives the similarity of region r to the visual term t, according to the low level feature of the lexicon of t, and *cover(r,I)* is the percentage of the area covered by r in I.

As we said before, *TF* just takes into considerations the content of a single image. However, the relevance of a term associated with an image should also take into consideration global aspects related to the dataset and the concepts to be recognized. In order to obtain the final weight of a term in an image we have compared two different techniques.

The first technique is the *TF*IDF* [12][13] technique widely used in text retrieval systems. It says that the weight of a term is directly proportional to the relevance of a term in the document (image in our case) and inversely proportional to the frequency of the term in the entire collection (Inverse Document Frequency, or simply *IDF*). We define the inverse document frequency of term t (IDF_t) as in traditional text retrieval systems:

$$IDF_t = \log_e \frac{N}{n_t}$$

The weight of a term t in an image I is therefore obtained as

$$w_t^I = TF_t^I * IDF_t$$

The second technique that we have considered uses the information gain (*IG*) [4], which is typically used for feature selection. Given a concept c_i that we want to recognize, the *IG* says that the term t is relevant when it is able to discriminate when a document contains c_i. The *IG* of a term t for a concept c_i can be expressed as

$$IG_{t_k,c_i} = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t,c) \log_2 \frac{P(t,c)}{P(t)P(c)}$$

where c_i indicates that a random document x belong to the category c_i and $\bar{c}_i$ indicates that it does not belong to the category c_i, t_k indicates that the term t_k occurs in x and $\bar{t}_k$ indicates that it does not occur in x. For example, $P(\bar{t}_k, c_i)$ indicates the probability that, for a random document x, term t_k does not occur in x and x belongs to category c_i.

The weight of a term t in image I, when we want to recognize the concept c_i, is therefore obtained as

$$w_t^{I,c_i} = TF_t^I * IG_{t_k,c_i}$$

The two weighting strategies above are compared in the experiment section.

4. CLASSIFICATION

We want to represent the content of images by means of a finite set of concepts, denoted by labels as, for instance, car, person, landscape, flower, city, people, countryside, sport, etc. In order to recognize these concepts in an image we build *binary classifiers*, that is, classifiers able to decide if a specific concept is present (recognized) or not in an image. In this respect, we need to build a specialized classifier for each concept that we want to identify. A classifier takes as input an image and it says if the associated concept is present or not in the image. In order to obtain a complete description of an image, all available classifiers should be applied to the image to be classified.

We build classifiers by applying the Support Vector Machine (SVM) [5] technology to the representation of images, in terms of vectors of weighted visual terms, described in previous sections. To execute the learning phase we have used the kernel adatron algorithm. It was first introduced in [2] as a perceptron-like procedure to classify data and then a kernel-based was proposed in [7]. In our implementation we used the SVM library in [3].

The performance of the SVM classification is strongly related to the choice of the kernel function, the kernel parameters, and some parameters related to the adatron algorithm, such as the penalty parameter *C*, the maximal tolerance on the margin, and maximum number of iterations. The kernel that we have chosen is the Gaussian Radial Basis Function (RBF) given its capability to recognize separate areas of the vector space where positively classified elements can be found. In order to automatically find the optimum values for the penalty parameter *C* and the variance σ of the RBF we have used the *v* cross-validation technique.

5. EXPERIMENTS

We used the ITI segmentation algorithm [10] to obtain the regions used for generating the visual lexicon and for indexing images of the experiments. The ITI algorithm was set to extract about 10 regions from each image. From each region we extracted the five MPEG-7 [11] visual descriptors (Scalable Color, Edge Histogram, Dominant Color, Region Shape, Homogenous Texture), by using the MPEG-7 reference software. We used a subset of 1000 randomly chosen images as training set to generate the visual lexicon. Various mono-feature visual lexicons were generated by using the 5 MPEG-7 visual descriptors. For each of the 5 features we used the clustering and the random choice techniques to identify visual terms, and we generated two different mono-feature visual lexicons of different sizes containing respectively 100 and 1000 visual terms. The proposed approach was tested by using each mono-feature visual lexicon and by combining 5 mono-feature visual lexicons, corresponding to the 5 MPEG-7 visual descriptors, in a single multi-feature lexicon.

To perform the experiments, we used a subset of the COREL collection, containing images belonging to 5 different categories (buses, roses, colleges, mountains, and deer). Each category is composed of 100 examples. Images within each category are randomly divided in 10 folders each with 10 elements to perform *v* cross-validation. We have tested one category at a time. Positive examples are taken from the 100 images composing the category itself, while negative examples are randomly chosen from the other categories.

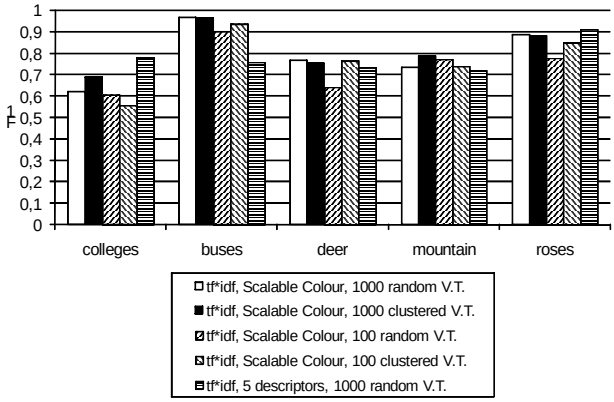

Figure 1: Classification performance using a visual lexicon obtained by random choice and k-medoid clustering algorithm. Sizes of the visual lexicon are 10 and 1000. We used a mono-feature visual lexicon obtained by using the MPEG-7 Scalable Color only. We also compare a multi-feature visual lexicon using all 5 MPEG-7 descriptors. All tests were obtained using tf*idf.

5.1. Results

Objective performance measurement was obtained by using the well known F1 measure defined as

$$\mathrm{F1} = \frac{2a}{2a+b+c},$$

where *a* are the positive examples correctly classified, *b* are the positive examples incorrectly classified, *c* are the negative examples incorrectly classified. We computed the F1 measures for each concept according to each tested setting.

The first test that we performed was intended to compare the performance obtained using various strategies for generating the visual lexicon. We compared visual lexicons of size 100 and 1000 elements, obtained using both the random and clustering visual term generation. We compared individually the 5 mono-feature visual lexicons, corresponding to the 5 MPEG-7 visual descriptors, using tf*idf to associate weight with images and visual terms. The best performance was obtained by using the Scalable Color visual descriptor with all tested settings and results are reported in Figure 1. For brevity, we do not report results obtained with all visual descriptors. F1 values were in general very high, reaching 0.95 in some cases (buses concept). From the histograms we can infer that there is not an evident difference in performance between the random and the clustering strategies and between the visual lexicons of size 100 and 1000. Consider that the cost of using the clustering strategy is *O*(*k***N***iter*), where *k* is the number of visual terms, *N* is the number of initial regions of the training set, and *iter* is the number of iterations of the k-medoid algorithm. On the other hand the cost of the random strategy is simply *O*(*k*), given that we have just to chose *k* random regions. The cost of the clustering strategy is much higher than that of the random strategy. However, clustering does not offer a performance that justifies its use. We can also observe that there is not a significant difference in performance when the size of the visual lexicon changes from 100 to1000. This suggests that the information, to classify images with this technique, does not need to be very fine grained. In Figure 1 we also report the results obtained by using a multi-feature visual lexicon. It can be seen that the results are comparable (sometimes a bit better sometime a bit worse) to those obtained when we used the best

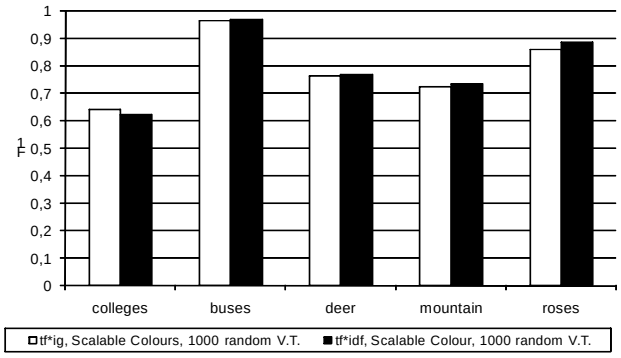

Figure 2: Classification performance using tf*ig and tf*idf, and lexicon obtained by random choice. We just compared a mono-feature visual lexicon obtained using the MPEG-7 Scalable Colors descriptor

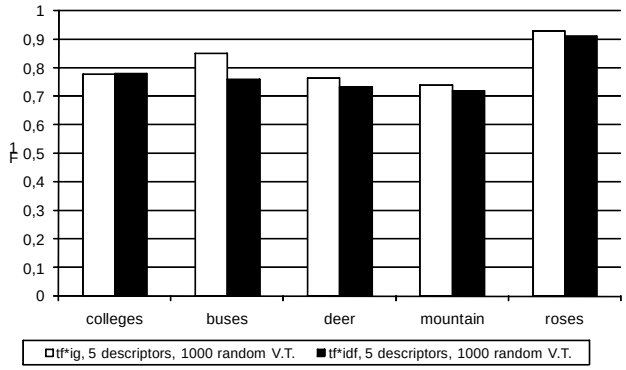

Figure 3: Classification performance using tf*ig and tf*idf, and visual lexicon obtained using random choice. We used a multi-feature visual lexicon containing all 5 MPEG-7 descriptors.

mono-feature visual lexicon, which is the one obtained using Scalable Color. Note that in general, it is not possible to know in advance what the best mono-feature visual lexicon is. The performance might depend, on the test sets, the training sets, the concepts. The advantage of the use of the multi-feature lexicon is that we do not have to decide in advance what the best visual descriptor is. In fact, thanks to the weighting strategy, the method offers good performance (comparable to the best visual descriptor) in all circumstances.

In the second test we compared the performance obtained using the tf*idf and tf*ig weighting strategies. Also in this case we used a mono-feature lexicon of size 1000 obtained using the Scalable Color visual descriptor with the random strategy. Surprisingly, our tests showed that there is not a significant difference in performance between the two methods. Results are reported in Figure 2.

For completeness we ran the same test using the multi-feature visual lexicon. Results are reported in Figure 3. In this case the difference in performance of the two weighting strategies becomes more evident. The tf*ig method returns the best results on average. We believe that this effect is due to the capability of the tf*ig method to better select the relevant terms for a concept, which is more important in a multi-feature lexicon, where selecting a term has also the effect of deciding the importance of a visual descriptor with respect to the other. The effect was less evident in the previous experiment given that we tested the mono-feature visual lexicon obtained using the Scalable Color visual descriptor, which, as we said before, is the best descriptor for the dataset and concepts that we used.

6. CONCLUSIONS

We presented a technique for automatic recognition of image content based on machine learning techniques. A classifier has to be built for each concept that we want to recognize. We use an image visual representation based on visual terms representing regions of images. Experiments provided evidence that generating the visual terms by simply choosing a number of random regions, from the set of regions extracted from a training set of images, offers a performance comparable to that of selecting visual terms by using a clustering techniques. Clearly the cost of randomly selecting the images has a cost negligible, compared to that of the clustering algorithm. In addition, we have evidence that the weighting techniques based on the information gain, in conjunction to the use of several different types of visual descriptors, to describe the visual appearance of regions, beats weighting based on tf*idf.

7. REFERENCES

[1] G Amato, V. Magionami, P. Savino, Image Indexing and Retrieval Using Visual Terms and Text-Like Weighting, Proceedings of the *DELOS Conference on Digital Libraries*, Tirrenia (PI), Italy, 13-14 February 2006

[2] J. Anlauf and M. Biehl. The adatron-an adaptive perceptron algorithm. *Europhysics Letters*, vol.10:pp.687–692, 1989.

[3] G. Caron, SVM in Java, http://www.site.uottawa.ca/~gcaron/svm.htm

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, NY, USA, 1991.

[5] N. Cristianini and J.Shawe-Taylor, *An Introduction to Support Vector Machines, and other kernel-based learning methods*, Cambridge University Press, Cambridge, UK, 2000

[6] Julien Fauqueur, Nozha Boujemaa, Mental Image Search by Mental Composition of Boolean Categories, *Multimedia Tools and Applications*, Volume 31, Number 1, October 2006.

[7] T.-T. Frieß, N. Cristianini, and C. Campbell. The Kernel-Adatron algorithm: a fast and simple learning procedure for Support Vector machines. In *Proc. 15th International Conf. on Machine Learning*, pages 188–196. Morgan Kaufmann, San Francisco, CA, 1998.

[8] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York, 1990.

[9] B. Le Saux, G. Amato, Image Classifier for scene analysis, *ICCVG 04, International Conference on Computer Vision and Graphics*, Warsaw, Poland September 22-24, 2004

[10] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. Still image segmentation tools for object-based multimedia applications. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(4):701–725, June 2004.

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

[12] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, 1983.

[13] I. H. Witten, A. Moffat, and T. C. Bell. Bell: Managing Gigabytes: Compressing and Indexing *Documents and Images*. Morgan Kaufmann, 1999.