IMAGE CLASSIFIERS FOR SCENE ANALYSIS

Bertrand Le Saux and Giuseppe Amato
ISTI - CNR di Pisa *
Via G. Moruzzi, 1 - 56124 - Pisa - Italy
bertrand.lesaux@isti.cnr.it, giuseppe.amato@isti.cnr.it

Abstract The semantic interpretation of natural scenes, generally so obvious and effortless for humans, still remains a challenge in computer vision. We intend to design classifiers able to annotate images with keywords. Firstly, we propose an image representation appropriate for scene description: images are segmented into regions and indexed according to the presence of given region types. Secondly, 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.

Keywords: scene analysis, feature selection, image classification, kernel methods

Introduction

How might one construct computer programmes in order to understand the content of scenes ? Several approaches have already been proposed to analyse and classify pictures, by using support-vector machines (SVM) on image histograms [3] or hidden Markov models on multi-resolution features [10]. To capture information specific to an image part or an object, approaches using blobs to focus on local characteristics have also been propounded [6].

We believe that a segmentation of images into regions can provide more semantic information than the usual global image features. The images that contain the same region types are likely to be associated with the same semantic concept. Hence, one has to design a classification scheme to test the co-presence of these region types.

This paper is organised as follows: § 1 explains how the scene information is represented by the means of *presence vectors*. The scene classifiers are described in § 2, and § 3 evaluates the proposed method.

*This research was partially funded by the ECD project and the Delos Network of Excellence

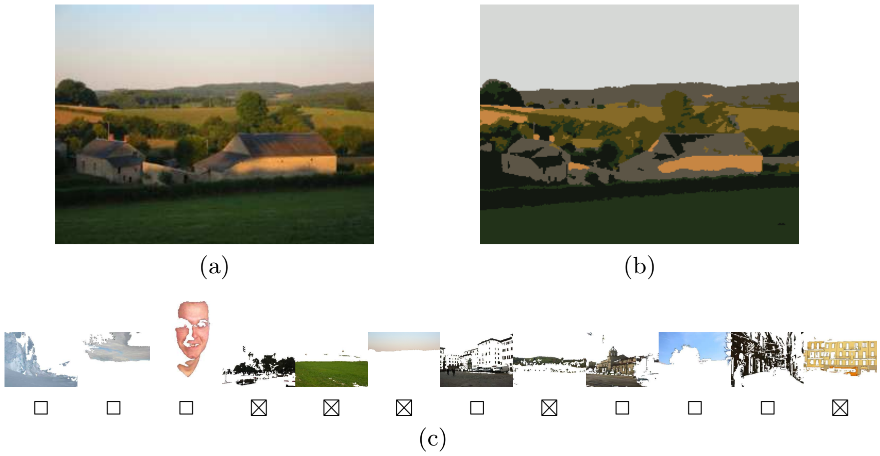

Figure 1. Feature extraction: the original image (a) is first segmented (b) then the boolean presence vector (c) is extracted by comparing the image regions to those in the region lexicon obtained with clustering techniques.

1. Feature extraction

1.1 Collecting region types

The *region lexicon* - the range of the possible region types that occur in a set of images - is estimated through the following steps:

- a training set of generic images provides a range of the possible real scenes;
- each one is segmented into regions by using the mean-shift algorithm [4];
- the regions are then pulled together and indexed using standard features - mean colour, colour histogram - for visual description;
- finally these index are clusterised by using the Fuzzy C-Means algorithm [2] to obtain categories of visually-similar image regions: each cluster represents a region type.

1.2 Representing the content of images

Given a region lexicon, 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 its similarity - in the visual-feature space - to the regions of the image. For instance, it is likely that a *countryside* image will contain sky, greenery and dark ground regions (cf. figure 1).

2. Automated image classification

We aim to annotate images with keywords. The keyword is basically a mnemonic representation of a concept such as people, countryside, etc. We define a binary classifier for each considered concept through a two-step process. Firstly a feature selection allows to determine which region types are meaningful to recognise a concept. Secondly a kernel classifier is used to learn a decision rule based on the selected region types.

2.1 Feature selection

In our application, the feature selection (FS) is a filtering phase [9]. The most standard way to select features consists in ranking them according to their individual predictive power.

Let Y denote a boolean random variable for the scene keyword to associate with the image. We denote $F_1, \ldots, F_p$ the boolean random variables associated with each feature, i.e. region type.

Information theory [8] provides tools to choose the relevant 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 $H(Y) = -\sum_y P(Y=y)\log(P(Y=y))$. 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:

$$I(Y, F_j) = H(Y) - H(Y|F_j) \tag{1}$$

The probabilities are estimated empirically on the training samples as the ratio of relevant items to the number of samples. The selected features are the ones which convey the largest information $I(Y, F_j)$ about the class to predict.

2.2 Kernel-adatron classifiers

The adatron was first introduced as a perceptron-like procedure to classify data [1]. A kernel-based version was then proposed [7]. It solves the margin-maximisation problem of the SVM [11] by performing a gradient ascent.

The training data-set is denoted $T = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Each x_i is a reduced presence vector (with only the selected features) and y_i

Table 1. Error rates for various keywords: comparison of various classification schemes applied to presence vectors. Null errors on the training set (denoted as "train error") indicate the classifier over-fits to the data. It is better to allow a few mis-classified training samples in order to obtain better results on the test set and thus insure a good generalisation.

keyword	linear adatron		kernel adatron		FS + kernel adatron	
	train error	test error	train error	test error	train 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 %

is true or false depending on the fact that the image is - or is not - an example of the concept to learn. For a chosen kernel K, the algorithm estimates the parameters α_i and b of the decision rule, which tests if an unknown presence vector x corresponds to the same concept:

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

3. Experiments

3.1 Data-set

The data-set is composed of 4 classes of images containing 30 instances of a particular scene: *snowy, countryside, streets* and *people* and of a fifth one consisting of various images used to catch a glimpse of the possible real scenes. In the experiments, error rates are averaged on 50 runs using cross-validation with training of the classifier on 80% of the set [5].

3.2 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 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*. Finally, by selecting the most informative region types for each category ("FS + kernel adatron"), we can obtain performances on complex scenes as good as those on simple ones.

Figure 2. The meaningful regions correspond to the region types that have a high mutual information with the label to predict. The upper row shows the original images and the lower row shows only the meaningful regions in these images.

3.3 Meaningful regions

For each concept, the feature selection allows to retrieve the meaningful parts of the image: only these regions are then used by the classifier to recognise a given keyword. Figure 2 shows that the selected region types are consistent with what was intuitively expected: green ones are used to recognise *countryside*, skin-coloured ones *people* and white ones *snowy* landscapes.

3.4 Evaluation

Our approach is compared with a SVM applied to image histograms [3]. 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 colours. However, the performance is less effective for the more complex types of scene: various backgrounds make the generalisation 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 in regions and the feature selection allow to catch the details that permit to differentiate these images from others without over-fitting.

Conclusion

We have presented in this article a new approach to scene recognition that intends to identify the image-region types in a given scene. Both

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	
	train error	test error	train error	test error
snowy	0.0 %	5.0 %	2.4 %	8.5 %
countryside	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 %

the image representation and the classification scheme are particularly appropriate for scene description and provide a good trade-off between the classifier performance and the prevention of over-fitting. Moreover the method is robust, since it does not require a fine tuning of a complex algorithm but on the contrary uses a succession of simple procedures.

References

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

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

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

[4] D. Comaniciu and P. Meer. Robust analysis of feature spaces: Color image segmentation. In *Proc. of CVPR*, pages 750–755, San Juan, Porto Rico, 1997.

[5] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification.* Wiley, New-York, 2nd edition, 2000.

[6] P. Duygulu, K. Barnard, J.F.G. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. of ECCV*, volume 4, pages 97–112, Copenhagen, Denmark, 2002.

[7] T.-T. Friess, N. Christianini, and C. Campbell. The kernel-adatron algorithm: a fast and simple learning procedure for support vector machines. In *Proc. of ICML*, Madison, Wisconsin, 1998.

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

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

[10] 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.

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