Combining SVM and Graph Matching in a Bayesian Multiple Classifier System for Image Content Recognition

Bertrand Le Saux and Horst Bunke

Institut für Informatik und Angewandte Mathematik,
University of Bern, Neubrückstrasse, 10, CH-3012, Bern, Switzerland
{lesaux, bunke}@iam.unibern.ch

Abstract. In this paper, we propose an approach to image content recognition that exploits the benefits of different image representations to associate meaning with images. We choose classifiers based on global appearance, scene structure and region type occurrence, and define confidence measures on their output. The resulting posterior probabilities of the classifiers are combined in a Bayesian framework. We show that this method leads to a robust and efficient system that contributes to reducing the semantic gap between low level image features and higher level image descriptions.

1 Introduction

In recent years, many approaches to image classification and scene recognition have been proposed. They usually utilize binary classifiers that learn from a set of indexed images to recognize a given concept (such as an image with people, an outdoor image, etc.) and strongly depend on the image representation they use. Such programs satisfy needs in image retrieval and computer vision, and could possibly be applied to a wide range of areas including robotics, digital libraries, and web searching.

Image classification has been performed by using support vector machines (SVM) on image histograms [1] or more recently hidden Markov models on multi-resolution features that capture more information [2]. These methods do not take into account that human description of an image content is rarely global but often specific to an image part. Several approaches use flexible models to localize objects in the image and perform object class recognition [3]. For scene recognition, both background and foreground objects are important, and thus image segmentation has usually been the basis to include local information in the image representation. The resulting indexes were more and more meaning-oriented: from image blocks [4] and attributed relational graphs [5] to presence vectors that check the occurrence of given region types [6].

Simple models for image representation are efficient when they fit salient features of the class of scenes to recognize. But they can be totally useless in other cases. Thus more and more complex image representations were proposed. These methods rely on the capacity of powerful classifiers to separate different image categories. Unfortunately, these classifiers are prone to overfitting if not enough training data are available, as is often the case in practice. We show in this paper that a system of classifiers that use

simple image representations related to different semantic levels, from global features to region type, and finally to region relationship information, is more suitable to learn and recognize a concept. These classifiers can be combined efficiently using a Bayesian framework.

This paper is organized as follows. Section 2 presents the various classifiers we use and shows their respective advantages. Section 3 introduces the Bayesian framework that allows us to combine these classifiers. In Section 4, posterior probabilities are defined for the classifiers. We present experiments and discuss their results in Section 5. Finally, conclusions are drawn in Section 6.

2 Image Classifiers

In this section we present the different classifiers and the scene features they characterize. For every class we build a binary classifier on each of the three representations. In the recognition task, images are tested with classifiers corresponding to different concepts. As a result, every image can be assigned to no, one or more than one concept.

SVM on Color Histograms. In [1], images are represented by simple 16-bin histograms of the color distribution in the Luv color space. A SVM with a Gaussian kernel is used to learn a decision rule in this feature space. This classifier, however, is inappropriate as soon as the distinction between the classes depends on local image regions. In this case it tends to overfit on the zones of the images that have no discriminative power, like the background. However, when a type of scene has a strong density peak on some colors or is visually consistent over all images of a class, the simplicity of this model insures that it is particularly efficient.

k-NN on Image Graphs. While the image representation described in Section 2 conveys only low-level and global information about the image, our second representation gives insight about the structure of the image and thus provides semantic information at the medium level. In [7], images are represented as graphs of regions; nodes of the graph are region histograms and an edge exists between two nodes if the corresponding regions are adjacent. A graph-edit distance based on the A^* algorithm [8] is used to match image representations. In this approach, one needs to define a set of possible graph-edit operations and assign a cost to each of them. We use the distance between the color histograms that represent the regions as the vertex substitution cost. To make vertex deletion easier on large graphs, we define its cost as the inverse of the number of vertices. To have comparable costs, this is extended to all deletion and insertion costs. Based on the graph-edit distance, a k-Nearest-Neighbor (k-NN) procedure classifies the images.

The semantic information provided by this classifier is twofold: first, it encodes the proximity of the regions in the image and second, the use of a graph-edit distance allows one to implicitly build a model of the class to learn by selecting the more similar regions between two images.

SVM on Presence Vectors. In this approach, images are indexed by boolean vectors that indicate the presence of particular region types, such as greenery, sky, or skin, chosen from a 50-region lexicon [6]. This lexicon is built by unsupervised clustering

of a set of regions based on color. The same training set is used to build the lexicon and train the classifiers and thus no manual intervention is needed to index regions nor images. A decision rule based on the resulting presence vectors is then evaluated by means of a SVM with a polynomial kernel. In the linear case, this approach checks the co-occurrence of region types. This image classifier is efficient when some region types are particularly salient, eg. skin regions to recognize images with people. But it depends on the quality of the region lexicon and might suffer from region types that have been badly estimated.

This kind of classifier offers even more semantic information than the previous ones since the learned decision rules make the composition of image classes explicit. For example, a “street” picture usually contains “building” and “road” region types, but no “field” or “snow” ones.

3 Bayesian Combination

Multiple Classifier Systems (MCS) have successfully been used to solve difficult classification problems [9]. In this paper we present a MCS based on the three classifiers described in Section 2. This section presents various combination rules based on a Bayesian framework [10,11].

Let $\mathcal{I}$ denote the set of images, and X a random variable on $\mathcal{I}$ standing for the distribution of images. We denote by Y a boolean random variable for the class to predict, i.e. the type of scene to associate with the image. Let us assume that we have R image representations denoted as $x_1, \dots, x_R$, that are modeled as functions $f : \mathcal{I} \rightarrow \mathcal{F}_k$ where $\mathcal{F}_k$ is the feature space associated with x_i . These representations correspond to the input of each classifier. We denote by $X_1 = x_1(X), \dots, X_R = x_R(X)$ the random variables associated with these representations.

From a Bayesian point of view, an image x should be assigned to the class with the maximum probability according to all classifiers, i.e x belongs to the class to predict if:

$$P(Y = 1|x_1, \dots, x_R) \geq P(Y = 0|x_1, \dots, x_R) \quad (1)$$

Practically, we can have only some estimates of the individual posterior probabilities $P(Y = 1|x_i)$ of each classifier i . Combination rules combine these estimates to approximate $P(Y = 1|x_1, \dots, x_R)$ in the best possible way.

3.1 Mean Rule

If we use individual classifiers that are good enough, we can assume that they are wrong on a few data samples only and that their answer differs from the overall answer by a small classifier error with no bias. All estimates can be used to obtain a better overall estimate by computing their average. From (1), we get the mean rule:

$$\frac{1}{R} \sum_{i=1}^R P(Y = 1|x_i) \geq \frac{1}{R} \sum_{i=1}^R P(Y = 0|x_i) \quad (2)$$

If there are outliers among the posterior probabilities, a robust estimate of the average like the median can be used to prevent them from affecting the final decision. The median rule is:

$$\text{med}_{i=1}^R P(Y = 1|x_i) \geq \text{med}_{i=1}^R P(Y = 0|x_i) \quad (3)$$

3.2 Product Rule

Using Bayes rule, (1) is equivalent to:

$$\frac{P(x_1, \dots, x_R|Y = 1)P(Y = 1)}{P(x_1, \dots, x_R|Y = 0)P(Y = 0)} \geq 1 \quad (4)$$

If the features are different, we can assume that the feature spaces $\mathcal{F}_1, \dots, \mathcal{F}_R$ are conditionally statistically independent: $P(x_1, \dots, x_R|Y = y) = \prod_{i=1}^R P(x_i|Y = y)$. We obtain that:

$$P(Y = 1) \prod_{i=1}^R P(x_i|Y = 1) \geq P(Y = 0) \prod_{i=1}^R P(x_i|Y = 0) \quad (5)$$

or, by using once more Bayes rule, in terms of posterior probabilities:

$$\begin{aligned} (P(Y = 1))^{-(R-1)} \prod_{i=1}^R P(Y = 1|x_i) &\geq \\ (P(Y = 0))^{-(R-1)} \prod_{i=1}^R P(Y = 0|x_i) &\end{aligned} \quad (6)$$

Eq. (6) is known as the product rule. It is critical in the sense that a single classifier with an output close to 0 can heavily influence the result of the whole ensemble.

3.3 Sum Rule

If we assume that $P(Y = 1|x_i) = P(Y = 1)(1 + \delta_i)$, i.e that the posterior probabilities do not differ too much from the prior probabilities, then, by substituting it in the left term of (6) and keeping only the first-order terms, we obtain:

$$(P(Y = 1)) \prod_{i=1}^R (1 + \delta_i) = P(Y = 1) + \sum_{i=1}^R \delta_i \quad (7)$$

And finally by expressing δ_i according to the probabilities, we find the sum rule, that outputs the class label if:

$$\begin{aligned} (1 - R)P(Y = 1) + \sum_{i=1}^R P(Y = 1|x_i) &\geq \\ (1 - R)P(Y = 0) + \sum_{i=1}^R P(Y = 0|x_i) &\end{aligned} \quad (8)$$

3.4 Other Combination Strategies

For comparison purpose, we present here other rules that are commonly used to combine classifiers. The first two ones also use the posterior probabilities. The max rule checks if:

$$\begin{aligned} (1 - R)P(Y = 1) + R \max_{i=1}^R P(Y = 1|x_i) \geq \\ (1 - R)P(Y = 0) + R \max_{i=1}^R P(Y = 0|x_i) \end{aligned} \quad (9)$$

The min rule is:

$$\begin{aligned} (P(Y = 1))^{-(R-1)} \min_{i=1}^R P(Y = 1|x_i) \geq \\ (P(Y = 0))^{-(R-1)} \min_{i=1}^R P(Y = 0|x_i) \end{aligned} \quad (10)$$

By using the majority voting rule, the MCS simply takes the same decision as the majority of classifiers. Only their binary outputs Δ_i^y , $y \in [0, 1]$ are considered:

$$\sum_{i=1}^R \Delta_i^1 \geq \sum_{i=1}^R \Delta_i^0 \quad (11)$$

4 Classifier Outputs

Since the three classifiers we use have a simple binary output, we need to define how to estimate their posterior probabilities to assign a class in order to combine them in a system more accurately than the simple voting rule.

4.1 SVM Case

SVM methods used for classification [12] project input data to a space of large (or infinite) dimension and use a linear decision boundary to separate different classes. Data x_i are classified by using the signed normalized distance d_{x_i} to the separating hyperplane.

In our application, only data with positive value are classified as belonging to the considered category. Data x_i with the signed distance d_{x_i} in the $] -1; 1[$ interval fall within the margin and are considered ambiguous. We define an estimate of the posterior probability of the SVM-based classifiers from the output value d_{x_i} by the following linear mapping, which approximates the real unknown distribution:

$$P(Y = 1|x_i) = \begin{cases} 1 & \text{if } d_{x_i} \geq 1 \\ (d_{x_i} + 1)/2 & \text{if } -1 < d_{x_i} < 1 \\ 0 & \text{if } d_{x_i} \leq -1 \end{cases}$$

4.2 k-NN Case

For classifiers based on the k-NN procedure, the posterior probabilities can be estimated simply by counting the number of positive neighbors k^+ among all the considered neighbors k :

$$P(Y = 1|x_i) = \frac{k^+}{k} \quad (12)$$

We use $k = 15$ in the graph-based classifier. This ensures that the probability estimates have enough different values.

5 Experiments

5.1 Data Set

The data set is composed of 200 images collected from the web. Four classes, consisting of 40 images each, contain instances of a particular scene type: *snow landscape*, *countryside*, *streets* and *people*. A fifth one consists of various generic images aimed

Fig. 1. Typical results of the histogram classifier (1), image graph classifier (2) and the presence vector classifier (3). Left-hand side shows images that are correctly retrieved by the considered classifier and only by this one. Right-hand side shows images that are typically misclassified by using this image representation.

to cover the whole spectrum of possible real scenes and thus used as negative samples for the classifiers. In the experiments, a training set of 150 instances (30 per category) is extracted randomly from the data set. The remaining instances are used as a test set. Error rates are averaged on 25 runs of holdout cross-validation. Some examples of each class are shown in Figure 1.

5.2 Strong and Weak Points of the Classifiers

Figure 1 shows images that illustrate the particular behavior of each classifier. The left-hand side of this figure shows the true positives that are recognized only by this kind of classifier and not by the other two, while the right-hand side shows images that are not recognized, i.e. either false positives or false negatives.

The histogram classifier recognizes the image [1-a] on the basis of the color only, while the other classifiers do not find the structure of the mountain landscapes they have learned on other images of the category. Due to the global and continuous image representation, this classifier also has a better tolerance to luminosity shifts than individual region matching (cf. image [1-b]). On the contrary, it hardly learns concepts characterized by images that have flat histograms, so there are many false negatives for concepts like *streets* (cf. image [1-c]) or *people*.

The image graph classifier allows to retrieve the highly structured image [2-a] although colors are almost lost completely in shadow. This can turn into a drawback when the image segmentation into regions is not reliable: in the close-up portrait of image [2-b], the face was segmented into several artificial regions, resulting in a meaningless graph representation for which the classifier is not able to find a good match.

Previous classifiers fail on images [3-a] and [3-b], which are partly overexposed (and so have density peaks on white colors), contain some heterogeneous elements (like the house in the country image) and have unusual structures. The presence vector classifier that tests the occurrence of some meaningful region types proves to be better suited to recognize these images. But it can also fail to find a good balance between two opposite clues, like the grass area and the face regions of image [3-c].

5.3 Comparison of Different Combination Schemes

The first round of experiments intends to test the complementarity of the various classifiers. The use of a MCS is promising if for one classifier that misclassifies some data, there is another one that can correctly recognize it. The oracle rule checks if for each test data there is at least one classifier that has the right answer. Table 1 shows the test errors of the individual classifiers and the oracle answer. We conclude that, for all concepts, all images can be recognized by at least one classifier, depending of the features

Table 1. Individual classifier errors and oracle prediction

	presence-vectors	region-graphs	histograms	oracle
snow landscape	15.15%	11.15%	3.52%	0.0%
country	6.30%	13.45%	7.64%	0.0%
people	10.30%	12.24%	11.88%	0.0%
streets	9.70%	17.94%	13.45%	0.0%

Table 2. Various combinations for the multi classifier system

	product	sum	max	min	mean	median	voting
snow landscape	7.15%	7.15%	8.24%	7.63%	5.58%	5.82%	6.67%
country	4.73%	6.55%	6.67%	6.79%	4.61%	4.61%	5.09%
people	8.48%	11.03%	12.73%	9.45%	6.18%	7.03%	6.91%
streets	7.88%	10.06%	11.15%	8.97%	6.42%	6.75%	7.39%

Table 3. Overall performances

	presence-vectors	region-graphs	histograms	mean-rule	MCS
mean error	10.36%	13.70%	9.13%		5.70%
standard deviation	6.31%	5.16%	7.74%		1.40%

the image representation focuses on. This is an indication that the individual classifiers have a high degree of complementarity.

In Table 2, we test various schemes for classifier combination. The sum, max and min rules give the worst results and do not perform better than the individual classifiers. The sum rule relies on the assumption that the posterior probabilities do not differ too much from the prior. This is true when the classifier outputs are ambiguous, due to noisy image representations for example, but all the classifiers we use are not weak learners and give good results, different from the prior. The max and min rules might be considered as rough approximations of the sum rule that emphasizes one of the classifiers. Performance of the max rule is below the sum rule, while the min rule performs slightly better on classes *streets* and *people*.

Among the other combining schemes, both the median and voting rules can be considered as variations of the mean rule. The median is another estimate of the mean, but here it gives slightly less good results due to the small number of classifiers. The voting procedure is equivalent to a mean rule applied to the binary outputs of the classifiers: the better results of the mean rule validate the estimates of the posterior probabilities that we have defined.

Both the product rule and the mean rule (and its derivatives) can be considered as successful since they globally improve the classification rates. The product rule corresponds to the true Bayes formula under conditional independence assumption and thus should lead to more precise results. However by multiplying the posterior probabilities, it gives too much importance to the classifier that assigns low probabilities to a class. By contrast, the role of the mean rule is to reduce the effect of the errors of the individual classifiers. In this case it acts on both the classifier uncertainties and the error from the probability estimation, and thus achieves the best results.

5.4 Comparison of the MCS vs. the Individual Classifiers

The comparison of error rates shows that the MCS with the mean rule outperforms all the individual classifiers (cf. Tables 1 and 2). The histogram-based classifier does better for one class only, the *snow landscape* one. This is due to the fact that images from this category do not have much variance in their color distribution (it corresponds mainly to ice and snow regions).

Table 3 shows the error rates of the individual classifiers and of the MCS on average over different image categories. The MCS allows to reduce error rates by more than 3 percent. It does not suffer from the inadequacy of a particular image representation and thus is more robust. Moreover, the standard deviation is highly reduced, which means that the MCS is more general-purpose and able to deal with different kinds of visual concepts that can meet the needs of various users.

6 Conclusion

We have presented in this article a novel approach to scene recognition based on a multiple classifier system. It uses image classifiers that characterize different semantic features: global appearance, image structure and presence of meaningful region types. We define posterior probability estimates of the classifier outputs and combine them in order to maximize the posterior probability of the whole system.

This approach achieves very good performance without using a classification method that requires a fine tuning or complex image representation like multi-resolution features. On the contrary, we use simple individual classifiers and rely on their combination to arrive at the best decision. This makes it really robust and suitable for recognizing various kinds of scenes. Thus the combination of multiple classifiers, as proposed in this paper, can bring great benefits to image indexing and search in image databases.

References

1. Chapelle, O., Haffner, P., Vapnik, V.: Svms for histogram-based image classification. *IEEE Trans. on Neural Networks* **10** (1999) 1055–1065
2. Li, J., Wang, J.Z.: Automatic linguistic indexing of pictures by a statistic modeling approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **25** (2003) 1075–1088
3. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: *IEEE Conference on Computer Vision and Pattern Recognition*. (2003) 264–271
4. Minka, T., Picard, R.: Interactive learning using a society of models. *Pattern Recognition* **30** (1997) 565–581
5. Beretti, S., Del Bimbo, A., Vicario, E.: Efficient matching and indexing of graph models in content-based retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **23** (2001) 1089–1105
6. Le Saux, B., Amato, G.: Image recognition for digital libraries. In: *ACM Multimedia / International Workshop on Multimedia Information Retrieval*. (2004) 91–98
7. Le Saux, B., Bunke, H.: Feature selection for graph-based image classifiers. In: *IAPR Iberian Conference on Pattern Recognition and Image Analysis*. (2005) 147–154
8. Nilsson, N.J.: *Principles of Artificial Intelligence*. Tioga, Palo Alto, CA (1980)
9. Kittler, J., Roli, F., eds.: *Multiple Classifier Systems, First International Workshop, MCS 2000, Cagliari, Italy*, Springer Verlag (2000)
10. Kittler, J., Hatef, M., Duin, R.P., Matas, J.: On combining classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **20** (1998) 226–239
11. Tax, D.M., van Breukelen, M., Duin, R.P., Kittler, J.: Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition* **33** (2000) 1475–1485
12. Vapnik, V.: *The Nature of Statistical Learning Theory*. Springer Verlag, New-York (1995)