A progressive learning method for symbols recognition

Sabine Barrat
LORIA - Université Nancy 2
BP 239
54506 Vandoeuvre-les-Nancy Cedex, France
barrat@loria.fr

Salvatore Tabbone
LORIA - Université Nancy 2
BP 239
54506 Vandoeuvre-les-Nancy Cedex, France
tabbone@loria.fr

ABSTRACT

This paper deals with a progressive learning method for symbols recognition which improves its own recognition rate when new symbols are recognized in graphics documents. We propose a discriminant analysis method which provides allocation rules from learning samples with known classes. However a discriminant analysis method is efficient only if learning samples and data are defined in the same conditions but it is rare in real life. In order to overcome this problem, a conditional vector is added to each observation to take into account the parasitic effects between the data and the learning samples. We propose also an adaptation to consider the user feedback.

Categories and Subject Descriptors

I.5 [**Pattern recognition**]: Clustering

Keywords

Conditional discriminant analysis, symbol recognition

1. INTRODUCTION

Symbol recognition is a field within graphics recognition to which a lot of efforts have already been devoted. Several approaches are based on feature descriptors [1] and due to the structural aspects of some symbols graph matching techniques [5] are suited to symbols recognition. However current symbols recognition methods have good results when we want to recognize few different symbols with low noise and often disconnected from the graphics. In real life, we have to distinguish in large symbol databases hundreds of different symbols, often complex and embedded in graphics, and those methods provide weak results. For these reasons, the problem of symbol recognition is far to be solved since in many cases it is impossible to assume that symbols can be performed on clearly segmented instances, as symbols are very often connected to other graphics and/or associated with texts. The well-known paradox therefore appears : in order to correctly recognize the symbols, we should be able to segment the input data, and reciprocally to correctly segment them, we need the symbols to be recognized!

This in turn means that it is usually not possible to perform symbols recognition by simply assuming that a reliable segmentation process is available, that the symbols have been clearly extracted, normalized and noise free. Under these conditions to improve the recognition it is necessary to carry out learning methods. In this paper we do not consider structural approaches [6] but statistical methods. A lot of classification methods have been proposed and they can be divided into two classes [3]: supervised and unsupervised classification. We focus here on a supervised learning method. More precisely we consider the linear discriminant analysis because the method is simple and fast and can be adapted to the recognition of symbols. However, the drawbacks of the discriminating analysis are that it is based on some assumptions which are not always checked due to a large variability of the real data. That is, discriminant analysis methods are efficient only if training data and the others are defined in the same conditions but it is rare the case for the reasons specified above. This can lead to erroneous trend in the classification and to overcome this problem we use a recent approach called conditional discriminant analysis[2]. It is a modified analysis which improves the learning data by a suitable control of the possible trend.

The rest of this article is organized as follows. In next section we recall the discriminant analysis theory (section 2). Then we describe the conditional discriminant analysis process (section 3) and show how to adapt it to a symbol recognition process. Results on a large database are presented in section 4 and some conclu-

SAC'07 March 11-15, 2007, Seoul, Korea

sions and guidelines for future work are given in section 5.

2. DISCRIMINANT ANALYSIS

2.1 Definitions and notations

The discriminant analysis provides rules of decision starting from learning samples with known classes (supervised learning).

Let :

- X_j be a vector representative of an observation :

  $$X_j = {}^t(X_{1,j}, X_{2,j}, ..., X_{p,j}),$$

  where p is the number of the characteristic vector.
- $\overline{X_l}$ be the barycenter of the class l : mean vectors of each variable in the class l.
- W be the intraclass covariance matrix supposed identical in each class, of size $p \times p$, symmetrical and regular.

Let a new observation x (1 line, p columns). To be able to affect this new observation, the posterior probability should be maximized which is equal to minimize the following quantity :

$$||x - \overline{X_j}||^2_{W^{-1}} = {}^t(x - \overline{X_j})W^{-1}(x - \overline{X_j}). \qquad (1)$$

This method is reliable only if the conditions of measurements are invariant i.e. if the data are observed under the same conditions during and after the learning phase which is not always warranted. Often, the conditions of measurement depend of significant factors of variability, and unknown trends can appear in experimental conditions. In this case the learning does not succeed by determining well the class of an new observation and the learning is thus partial.

3. CONDITIONAL DISCRIMINANT ANALYSIS

To overcome this problem of trend factor, we are interested in a suitably modified analysis which completes the initial learning by a progressive control of conditions of use. In many real situations the learning samples and other data are not observed under the same conditions. In these cases, the measurements taken on the observations depend on factors of trends which would be advised to consider. The idea is to add to each observation X, the observation of a random vector Y, representative of trend due to the experimental conditions. Moreover, a descriptor only is not robust enough, it can't cover every types of noises. Add a vector to each observation will enable the method robust at more noises.

This approach called conditional discriminant analysis was first proposed by A. Baccini [2] in a different domain with classical statistical units.

3.1 Definitions and notations

Let us suppose that the estimates of the $\overline{X_j}, j \in 1, 2, ..., s$ and of W were made beforehand. Let Y be a matrix with n lines and h dimensions which takes into account the phenomenon of trend. We consider the following assumptions:

1. The intraclass mean of Y is $\overline{Y}$ (the empirical mean), whatever the class j.
2. The variance of $Z = {}^t(X, Y)$ is written :

   $$W_Z = \begin{bmatrix} W_X & W_{XY} \\ W_{YX} & W_Y \end{bmatrix}$$

   where $W_X = W$ is the empirical intraclass covariances of X, $W_{XY} = {}^tW_{YX}$ is the covariances matrix of X and Y. W_Y is the covariance matrix of Y and supposed regular.
3. The intraclass variance of Z is supposed identical for each class and follows a normal law.

To be able to consider the factors of trend, one will carry out a decisional discriminant analysis no longer on X, but on Z, vector with $p + h$ dimensions. Thus, the changes carried out are :

1. The matrix W is replaced by the matrix W_Z,
2. $\overline{X_j}$ are replaced by $\mu_j = {}^t(\overline{X_j}, \overline{Y})$.

Let us $C = W_X - W_{XY}W_{Y^{-1}}W_{YX}$ supposed regular. According to the principle of the usual decisional discriminant analysis, one must assign an new observation ${}^t(x, y)$ to the class j which minimizes the quantity:

$$\begin{bmatrix} {}^t(x - \mu_j) \\ {}^t(y - \mu_j) \end{bmatrix} W_{Z^{-1}}[(x - \mu_j)(y - \mu_j)]$$

what is equivalent to minimize the expression :

$$||(x - \overline{X_j}) - W_{XY}W_Y^{-1}(y - \overline{Y})||^2_{C_{-1}} \qquad (2)$$

In the metric W^{-1}, the new observation ${}^t(x, y)$ is assigned to the class j for which the following expression is minimal [2]:

$$||(x - \overline{X_j}) - W_{XY}W_Y^{-1}(y - \overline{Y})||^2_{C_{-1}} =$$

$${}^t((X - \overline{X_j}) - W_{XY}W_Y^{-1}(y - \overline{Y}))$$

$$(W_X - W_{XY}W_Y^{-1}W_{YX})^{-1}((X - \overline{X_j}) - W_{XY}W_Y^{-1}(y - \overline{Y}))$$

The significant point in this analysis is the correction of x. However, the replacement of W^{-1} by C^{-1} improves theoretically the analysis.

In fact, the traditional discriminant analysis is based on diagonalization of $B_XW_X^{-1}$ where B_X is the matrix of covariances between groups and W_X the matrix of covariance within groups. The conditional discriminant analysis is based on the diagonalization of $B_Z W_Z^{-1}$ because we apply traditional discriminant analysis on Z.

We give in appendix some indications of the demonstration for formula (4).

3.2 Parameters estimation

It is supposed that the estimates of the parameters W_X, $\overline{X_j}$ and $\overline{\overline{X}}$ are made beforehand from the learning samples as showed in §2.2.2.

We have :

$$W_{XY}^N = \mathbb{E}[(X - \overline{X_j})^t (Y - \overline{Y})] = \mathbb{E}[X^t (Y - \overline{Y})]$$

$$= \frac{1}{N} \sum_{i=1}^{N} X_i{}^t (Y_i - \overline{Y}^N), \tag{3}$$

where:

- N is the number of data for which Y is available. We recall that Y is measured on the learning samples and the other samples.
- $\overline{Y}^N$ is the empirical mean of Y_i.
- W_Y is defined by the empirical covariances matrix of Y on the whole observations where this variable is available.
- Using assumption 2 (§3.1) W_{XY} is computed by means of the whole data, even if we do not know the class of belonging for a new observation X.

Each time a new observation is added to the learning process the formula 5 must be reconsidered for all the previous data. In this perspective, for complexity considerations we define the following reccurence formula:

$$W_{XY}^{N+1} = \frac{N}{N+1}(W_{XY}^N + X_{N+1}(Y_{N+1} - \overline{Y}^N)).$$

4. CHOICE OF THE PARAMETERS

4.1 Vector X

So that the discriminating analysis is done in the best possible conditions, it is necessary as a preliminary to choose relevant variables. In our context, these variables can be obtained by using one or more descriptors which allow us to extract the quantitative variables from the symbols.

To generate relevant variables, which allow a good discrimination of the data, we should choose descriptors robust to the noise, the deformations, and if possible having properties of invariance to some geometrical transformations. Indeed, these properties of invariance will correctly classify the same symbol independent of its position and of its size in the document graphics.

For practical reason our choice was a descriptor defined in [7] and other descriptors can be used in a similar fashion[1] This descriptor is based on the transform of the Radon transform is the projection of an image in a particular plan. This projection has interesting geometrical properties which make it a good descriptor. According to these geometrical properties, a signature of the transform is created. This signature checks the properties of invariance to some geometrical transformations, such as the translation and it scaling (after normalization). On the other hand invariance with rotation is restored by cyclic permutation signature or directly starting from its Fourier transform. Table 1 shows examples of signatures for a symbol which is scaled and rotate. Thus, for our discriminant analysis, for each symbol of the learning sample and test, we compute its signature.

4.2 Conditional vector Y

Y constitutes the key point of the method. The choice of its components is essential for the success of the approach, and can be made in two ways [2]:

- One can initially carry out an ”external” choice in considering measurements of one or more indicators independent of the class to which the symbol belongs and most likely to be well correlated with the parasitic effects of one possible trend.
- One can also seek to carry out a ”internal” choice in suitably analyzing the data (during and afterward learning) in order to discover possible combinations measurements of X which seems most characteristic of one possible trend while being independent of the group.

These two types of construction of Y are not excluded, and some components of Y could be obtained in the first way and the others being obtained with the second.

Whatever the type of construction chosen, components must satisfy these two constraints, otherwise the success of the analysis could be compromised. Moreover the selected variables must to be:

1. Representative of parasitic effects of the analysis, factors of trends.
2. Independent of the class of the symbol considered.

Our choice is made on the first type. We need to find and to build Y which measures the parasitic effects of the analysis and which is not taken in account in

[1] We recall that the aim of the paper is not to underline a particular descriptor.

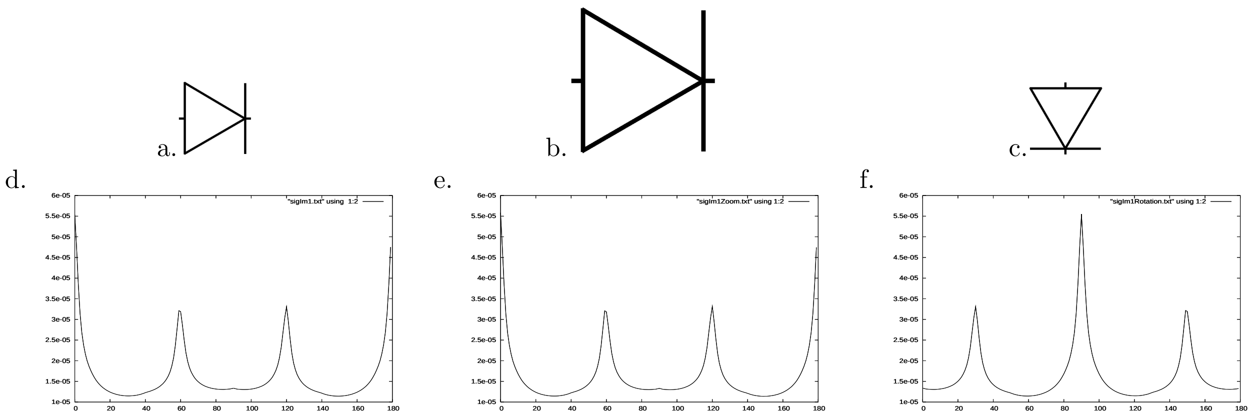

Table 1: Examples of signatures. a), b) and c) respectively a perfect symbol, the same symbol with a zoom ×2 and the perfect symbol turned of 90 degree. d), e) and f) the respective signatures.

the descriptors used for X. For these reasons, we minimize the deviations of the linear regression between the signature of the unknown symbol X and the signature of each model M representative of the different classes. That is,

$$\sum (X_i - \beta_0 - \beta_1 M_i)^2.$$

Thus, the selected Y describes the minimal linear deviations which are supposed to represent the additional information related to the degradation that a symbol should undergo in real conditions.

5. EXPERIMENTAL RESULTS

We used symbols from the GREC database for our tests [8]. This database (see fig 1) was created especially for symbols recognition contest.

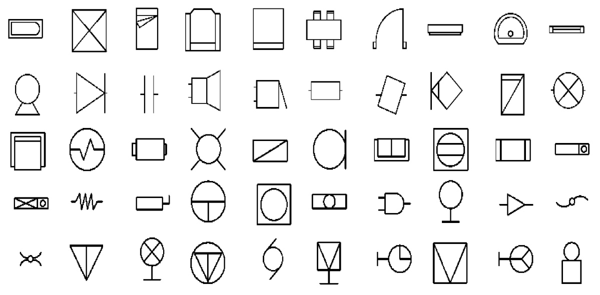

Figure 1: GREC database [8].

This database is mainly defined from two application domains, architecture and electronic, because the symbols are most largely used by teams and represent a great number of different forms. We have 50 different symbol models for which we applied some noises based on Kanungo [4] model. These noises are similar to noise obtained when a document is scanned, printed or photocopied. Thus we apply 4 kind of different noises, with different intensities, on each of the 50 models. Thus we get for each model 4 classes of noises. Each class of noises contains 100 noisy symbols but with different intensity. In this case we have a database with 20000 different symbols (4 classes of noises ×50 models ×100 intensities of degradation).

Next to simulate occlusion with dimensioning lines which often occur in document graphics we add random lines on each symbols of the database. More precisely, we create two new sets, each one containing 10000 symbols (one with 200 different black lines ×50 models, the other with 200 different white lines ×50 models). At all we have a database composed of 40000 symbols. For example, Table 2 presents different degradation applied on the same symbol model (TAB.2.a).

On this database we defined several tests of learning. We defined learning samples composed with two classes of noise and the recognition process is applied to the other classes with different noise. For example the learning samples is set to 5000 symbols : 100 degradations from the 50 models and the test samples is set to 10000 symbols belonging to the other classes. Then we calculated the recognition rate for this test samples with the discriminant analysis (DA), the conditional discriminant analysis (CDA) and the conditional discriminant analysis with user feedback. In the last case a user gives to the system his/her opinion (correctly or badly classified) at different moments (here every ten symbols) of the recognition process. This interactive procedure result increases recognition rate since according to the user opinion the learning is updated. Table 3 shows the results obtained with these tests. We can notice that from the beginning of the learning and until half, the DA and the CDA have similar behavior: the recognition rate decreases gradually from 92% to 74% for the DA and from 85% to 74% for the CDA. However the DA gives rise to slightly better results (a recognition rate approximately of 5 to 10 percent higher). Approx-

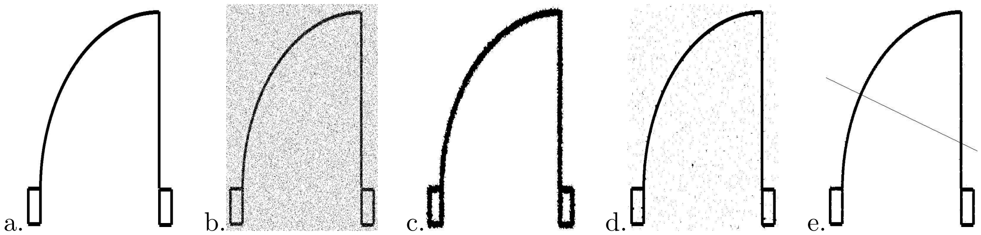

Table 2: Example of a symbol model a) on which we applied different degradations b), c), d) and e).

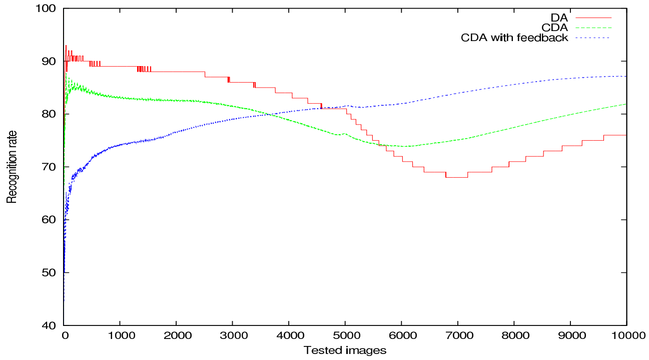

Table 3: Evolution of recognition rate of the CDA and the CDA with feedback compared to DA with a learning on 5000 symbols (100 symbols by class, 50 classes in all) and tests on 10000 (200 symbols by class, 50 classes in all).

imatively from 5000 tested symbols (which correspond to the size of learning databases), the recognition rate of the DA decreases until 7000 tested symbols and increases then slightly to reach at the end 76% recognition rate. On the contrary the CDA increases gradually until the end of the learning to reach the rate of 82% and is thus better than the DA in the second part of the learning. Using the user feedback makes the recognition rate increases gradually during all the learning from 65% to 87% recognition rate. Moreover, results of CDA with feedback are better than without, because feedback takes into account only correct affectations.

6. CONCLUSION AND PERSPECTIVES

In this paper we have proposed an original adaptation of a method of conditional discriminant analysis. The results show the robustness of the approach to the scale compared to the discriminant analysis. Our choice for the complementary variable Y has carried on the linear deviations between the perfect model and the symbol to be recognized. We see that this complementary variable takes into account the effects of trends related to the limits of a descriptor to a high number and variability of symbols and the method of discriminant analysis. Furthermore we have shown experimentally that the user feedback improves the learning. Future works will be dedicated to take into account other disturbances on symbols. We wish to consider nonlinear deviations in the determination of the variable Y and to combine several descriptors together in the recognition process.

7. REFERENCES

[1] S. Adam, J. Ogier, C. Cariou, R. Mullot, J. Labiche, and J. Gardes. Symbol and character recognition: application to engineering drawings. *International Journal on Document Analysis and Recognition*, 3(2), 2001.

[2] A. Baccini, H.Caussinus, and A. Ruiz-Gazen. Apprentissage progressif en analyse discriminante. *Revue de Statistique Appliquée*, 49, 2001.

[3] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical Pattern Recognition: A Review. *IEEE Transactions on PAMI*, 22(1):4–37, Jan. 2000.

[4] T. Kanungo, R. Haralick, H. Baird, W. Stuezle, and D. Madigan. A statistical, nonparametric methodology for document degradation model validation. *IEEE Transactions on PAMI*, 22(11), 2000.

[5] J. Llados, E. Marti, and J. Villanueva. Symbol recognition by error-tolerant subgraph matching between region adjancy graphs. *IEEE Transactions on PAMI*, 23(10), 2001.

[6] B. Messmer and H. Bunke. Automatic learning and recognition of graphical symbols in engineering drawings. In *Graphics Recognition – Algorithms and Applications*, volume 1072 of *lecture notes in computer science*. springer verlag, 1996.

[7] S. Tabbone, L. Wendling, and J. Salmon. A new shape descriptor defined on the radon transform, 2006. Computer Vision and Image Understanding, 102(1).

[8] E. Valveny and P. Dosch. Symbol recognition contest : A synthesis. In *Graphics Recognition – Algorithms and Applications*, volume 3088 of *lecture notes in computer science*. 2004.