A SEGMENTATION FREE APPROACH TO SYMBOL EXTRACTION AND RECOGNITION FROM IMAGE DOCUMENT

Maurice Milgram,
Mattieu Jobert,
Bertrand Lamy.

Univ. Pierre & Marie Curie
Lab. of Robotic of Paris
4 Place Jussieu
75005 Paris FRANCE

Abstract:
We present a symbol recognition method without segmentation of the document. Our approach uses Zernike moments for the coding and a multilayered Perceptron for the classifier. Recognition is independant of position and scale of symbols.

1) Introduction

An important problem in document (map,schema,text,..) analysis is the automatic extraction and recognition of a symbol regardless of its position, size, and orientation. The current approaches to invariant two-dimensional shape recognition include statistical or global methods (moments, Fourier descriptors[2], slope density function, ..) and structural methods [4](grammars, automata, trees, graphs).

Our goal is to detect and to recognize symbols in documents like maps, schemas, text with figures. We have to cope with following problems:

- position, size and orientation are not known
- distorsions are possible
- shape/background separation

We want too that our system can be easily trainable to work on new symbols chosen by the users.

To fullfill all these constraints, we have designed a connectionist system working on a vector of features extracted from the image without msegmentation.

2)Feature extraction

In [1], Zernike introduced a set of complex polynomials wich form a complete orthogonal set over functions that are zero on the outside of the unit disc.

These functions are defined by:

Vn,m(x,y)=Rn,m(r).exp(imt)
n is a nonnegative integer
m is an integer with m <n n- m even
(r,t) are polar coordinates of (x,y)
Rn,m is the radial polynom:

$$Rn,m(r)= \sum_{s=0}^{\frac{n-|m|}{2}} \frac{(-1)^s.(n-s)!.r^{n-2s}}{s!(\frac{n+|m|}{2}-s)!(\frac{n-|m|}{2}+s)!}$$

Functions Vn,m are orthogonals , thus it is easy to reconctructed a grey level function from its Zernike moments given by:

$$An,m=\frac{(n+1)}{\pi}.\int\int_{x^2+y^2<1} f(x,y)V^*(r,t)dxdy$$

An image is defined as a grey level function on the unit disc and can be decomposed on this basis and can be expressed with a set of Zernike coefficients. Comparing Zernike approach to the classical moment approach is of course necessary. We see that Zernike coefficients have a magnitude invariant under rotation whereas moments need a normalization that is known to be very unstable, depending on the acuracy of the principal axis of the shape.

The drawback is that we have to transform Zernike coefficients to obtain translation and size invariance but this transformation is stable.

The choice of the Zernike method is also made to avoid edge following or skeletonization (like in Fourier descriptors or structural approaches).

3) Classifier

3-1:Architecture

We have chosen a neural classifier [3] with one or two hidden layers. Each output unit is connected to a specific set of hidden units of the last hidden layer (about 5 to 10). All units of the first hidden layer are connected to all input units.Hidden layers are fully connected. Activity of an input unit represents the value of a feature (about 30 features) extracted from Zernike coefficients of the image.

3-2:Learning algorithm

The learning algorithm uses the classical backpropagation of the error gradient with some improvements.

The backpropagation learning procedure [5] is a generalization of the least squares procedure that works for networks which have layers of hidden units. These networks can compute much more sophisticated functions than networks without hidden units like Rosenblatt's Perceptron. The learning is generally slow and lot of improvements have been proposed to accelerate this process.

The learning process is as follows: during a cycle (called an iteration), we present all prototypes (examples and counter examples) and compute the gradient for each weight of the error function . The error function is the quadratic mean of the difference between actual output of the network and the desired output. At the end of each cycle, we modify all weights according to the sum of these gradients.

A viscosity term is introduced to accelerate ravine followings and a pertubation procedure is used to escape from small local minimums.

Despite these improvements, the convergence rate remains very slow and we have successfully tried a new strategy for the presentation of prototypes. Instead of presenting all prototypes during each cycle, we present only a subset of protos. This subset is built at each iteration with randomly choosen protos. These protos are selected at random with a probability computed for each protos with the last error value for this protos. This probability is merely half the last error so, for a missclassified proto the error is high and so the presentation is quite sure. For a proto with a good answer of the network, there will seldom be a presentation. This strategy uses the redundancy of the set of protos and the acceleration factor depends on this redundancy. For some training set, the acceleration factor have been about ten.

The choice of examples and counter examples is crucial. It is clear that it is impossible to "fill" the N-space with only a few points and that the number of points is growing exponentially if we want to maintain a minimum density for each coordinate. For each letter, the number of examples was 45 and the number of counterexamples 41.

3-3:Extraction/recognition

The extraction/recognition process consists of scanning all the image with a circular mask. For each position, we first decide if there is a possibility to find a symbol. This decision is based on specific knowledge depending on the document we are working on. If the decision is positive, we compute the feature vector and the response of the neural classifier. At this stage, the user can define several decision rules.

The most popular being: "the winner takes all" ;this one is used on purpose when the user's aim is to force a decision. An other rule is to suspend any decision when the entropy of the response is too high.

4) Results

A lot of experiments have been done on several types of symbols like letters & signs. The recognition rate is very good but we have to be very careful whith it.

To make the presentation clearer, we have just reported results concerning the recognition of the letter "a", printed in lowercase.We have mainly used 23 Zernike coefficients which is a good compromise between complexity and accuracy. Networks are described by their number of cells for each layer. For instance, 24/10/1 corresponds to a network with 3 layers, the input layer being 24 cells, the hidden layer 10 cells and the output layer only 1 cell. The latter gives the degree of membership of the input for the "set of lowercase printed a".

Errors=(E1 E2) means that:

- average of an error on a training sample is E1
- maximum error is E2

The correct answer is always -1 or +1 so the error range is [0 2].

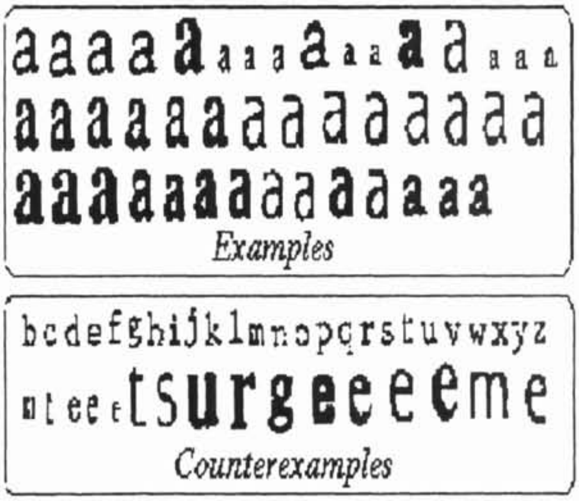
1) Three layers networks

```
 24/10/1 ---> Errors=(0.010 0.067)
#Iterations=33631
 24/15/1 ---> Errors=(0.012 0.063)
#Iterations=27710
 24/5/1 --->  Errors=(0.014 0.098)
#Iterations=27710
 24/20/1 ---> Errors=(0.011 0.072)
#Iterations=27700
```

2) Four layers networks

```
 24/10/5/1 ---> Errors=(0.005 0.030)
#Iterations=27700
 24/15/5/1 ---> Errors=(0.007 0.052)
#Iterations=27700
```

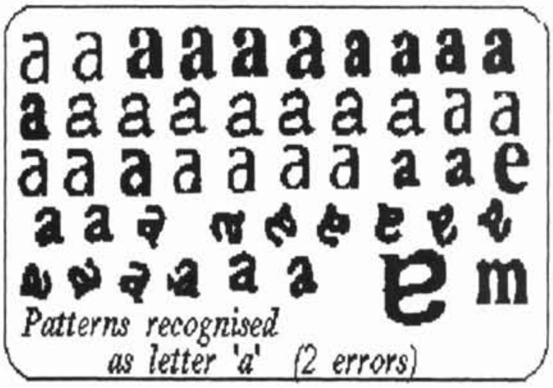
TRAINING SET



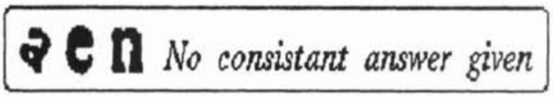
Examples

Counterexamples

RESULTS (3-LAYER NET)



Patterns recognised as letter 'a' (2 errors)



No consistant answer given

References:

[1] F.Zernike: Physica,vol 1,p689,1934

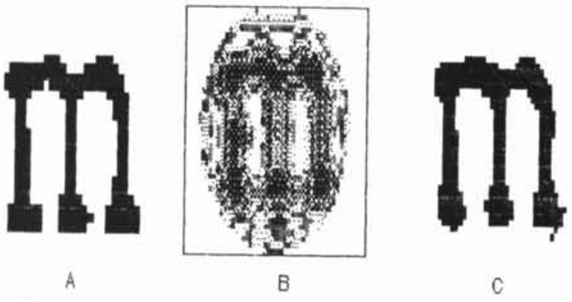
[2] Y.N.Hsu et al: "Rotational invariant digital pattern recognition using circular harmonic expansion" Appl. Opt. vol 21,pp 4012-4015,1982

[3] Rumelhart, McClelland: "Parallel Distributed Processing" vol 1; MIT Press

[4] K.S.Fu, "Syntactic Pattern Recognition and Application" Englewood Cliffs, NJ, Prentice Hall, 1982.

[5] Y. Le Cun, "Modèles Connexionnistes de l'Apprentissage" PhD Thesis, University Pierre et Marie Curie, Paris, France, 1987.

Zernike Moments - 12th Order



A - Original image
B - Gray level reconstructed image
C - Reconstructed image after histogram equalization and thresholding at 128