Distance Based Strategy for Supervised Document Image Classification

Fabien Carmagnac, Pierre Héroux, Eric Trupin

Distance Based Strategy for Supervised Document Image Classification

Fabien Carmagnac[1,2], Pierre Héroux[2], and Éric Trupin[2]

[1] A2IA SA
40 bis rue Fabert
75 007 Paris cedex - France
Fabien.Carmagnac@a2ia.com
WWW home page: http://www.a2ia.com
[2] Laboratoire PSI
CNRS FRE 2645 - Université de Rouen
76 821 Mont-Saint-Aignan cedex - France
WWW home page: http://www.univ-rouen.fr/psi

Abstract. This paper deals with supervised document image classification. An original distance based strategy allows automatic feature selection. The computation of a distance between an image to be classified and a class representative (point of view) allows to estimate a membership function for all classes. The choice of the best point of view performs the feature selection. This idea is used by an algorithm which iteratively filters the list of candidate classes. The training phase is performed by computing the distances between every class. Each iteration of the classification algorithm computes the distance d between the image to be classified and the chosen representative. The classes whose distance with this point of view differs from d are deleted in the list of candidate classes. This strategy is implemented as a module of A2IA FieldReader to identify the class of the processed document. Experimental results are presented and compared with results given by a knn classifier.

1 Introduction

Since a few years, handwritten recognition systems have improved (automatic zip code, form field or cheque amount reader) leading to commercial applications. The good reading rates of such systems allow to process less constrained documents (eg. order forms, invoices). To validate the reading results on such semi-structured documents, it is necessary to associate a reading model to each document class. This reading model contains information about the fields of the document: position, nature (letters, digits, consistency rules, meaning).

Some systems process only a unique kind of document. Others receive an heterogeneous stream of documents. In that case the system has to identify the document class (and therefore the corresponding reading model) of a given document, which means a document class identification module precedes the reading module. In many systems, the classes to discriminate are known before

the classification module conception [1] [2]. During the desing it is possible to determine the most discriminating features by studying the images representing the classes. On the other hand, systems are now conceived without knowing the number or the nature of the document classes. Then the classification module has to automatically determine the best features set for the given set of document classes during a learning step.

This paper presents a document classification module, which automatically performs feature selection on any set of document classes. This module is designed to be added to an existing system that reads handwritten fields on documents from a unique class. This system needs to know the reading model. By associating a reading model to each document class, the document classification module allows the system to process an heterogeneous stream of documents.

The system is intended to be sold to final users such as government services or banks, which are not specialists in document analysis nor classification specialists. Therefore the module determines reflexively the most discriminating set of features given a specific problem. If the available features set is not able to efficiently separate the images of a particular project, new features can be easily inserted inside the module.

Our strategy simultaneously performs the feature selection for a given problem and the document classification. It is based on the computation of distance between documents. In section 2, we remind the context in which the classification module is inserted and we expose the constraints attached to a commercial use. Section 3 introduces the main definitions. It details the notion of document class, presents the structure of the features used to extract feature sets, and finally introduces the concept of distance between document images according to to a feature and its metric. Section 4 details the principles of the classification process which is presented in section 5 and 6. Section 5 presents the different steps of the supervised learning phase. Section 6 describes the processes applied to document image to determine its class. Finally, experimental results are presented in section 7. These are compared with results given by a *knn* classifier.

2 Context and Constraints

We remind that the module presented in this article is added to an existing system that reads handwritten fields of forms from a unique class. The purpose of the module is to identify the class of the document so that the system can process documents from multiple classes. Once the document class is determined, the document is processed as done before by using the associated reading model. This reading model includes rules (location, nature, syntax, meaning, consistency) which allow to improve the automatic reading.

The use of the classification module as a part of a commercial product implies to respect some constraints. It must be able to discriminate the class of a document among a hundred classes, and the processing time has to be lower than a limit fixed by the final user (1 or 2 seconds per document with a 1GHz PC). This implies that the processing time is a sub linear function of the classes number.

The time needed to classify a document among 100 classes must not be 10 times greater than the time needed to classify a document among 10 classes. On the other hand, final users are not specialists in document analysis. They do not know how to determine the best parameters for the classification. The classification module must be able to automatically adapt its parameters to be as robust as possible on any document set, by evaluating its discrimination performance.

3 Definitions

This section defines the most important terms used in the description of the classification module principle. More accurately, the concepts of document classes, feature, and distance between documents are introduced and detailed.

As mentioned above, we define a document image class as a set of document images on which the same reading model is applied. In our case, the reading model is defined by the final user. A document class is then defined because of the subsequent processing. Two documents are in the same class if the fields to be read are at the same location even if the images look structurally very different. On the other hand, two documents which images are very similar can be in two different classes if, for example, two fields have their location inverted. For a given problem, two documents from the same class can be in two different classes in an other problem, if their reading model becomes different.

Notation 1 *Document image and document images set*
The document image classes set is called $ClassSet$.
The document images set is called $ImageSet$. *The set of images belonging to class* C *is called* $ImageSet_C$.

The features used can be numerical, syntactical and/or structural. It is possible to associate different metrics to each feature. Each metric allows to compute a distance between two document images.

Definition 1. *Feature and Metric*
The feature space associated to the feature F *is called* $Space_F$. *A metric associated to* F *is called* M_F.

$$F : ImageSet \rightarrow Space_F$$
$$M_F : Space_F \times Space_F \rightarrow [0, 1]$$

The features set is called $FeatureSet$. *The metric set is called* $MetricSet$.

The metric associated to each feature can be computed with Euclidean, Hamming, Max, Min [3], edition distance, graph distance.

Definition 2. *Distance between document images*
The distance between document images I_1 *and* I_2 *according the feature* F *is defined by :*

$$D_F : ImageSet \times ImageSet \rightarrow [0, 1]$$
$$D_F(I_1, I_2) = M_F(F(I_1), F(I_2))$$

A distance computed between two documents from the same class is called an intra-class distance. Otherwise, it is called an inter-class distance. For each metric M_F and for each class C, an image is chosen as representative. This choice is detailed in section 5.

Notation 2 *The image which represents the class C according to M_F is called $I^*_{M_F,C}$.*

The distance between an image I and a class C according to M_F is defined as the distance between I and $I^*_{M_F,C}$.

Definition 3.

$$\begin{aligned} D_{(M_F,C)} &: ImageSet \rightarrow [0,1] \\ D_{(M_F,C)}(I) &= D_{(M_F,C)}\left(I^*_{M_F,C}, I\right) \\ &= M_F(F(I^*_{M_F,C}), F(I)) \end{aligned}$$

Notation 3 *intra-class and inter-class distances*
The set of inter-class distances between $I^*_{M_F,C}$ and images from C' is called $DX_{M_F,C}(C')$.
The set of intra-class distances $DX_{M_F,C}(C)$ is called $DI_{M_F,C}$.

4 Classification principle

In classical approaches, a document image is represented by a point in a feature space. Then, classes correspond to clusters. The classification process consists in finding frontiers between these clusters. Our approach differs from commonly used classifiers. The feature space is projected in a one dimensional distance space. A point in this space represents the distance according a metric between a point of the feature space and a point of view. The representative $I^*_{M_F,C}$ defines a point of view. Thanks to the metric M_F, it is possible to compute the distances between $I^*_{M_F,C}$ and each other class.

Let $\tilde{I}$ be the image to classify and $\tilde{C}$ be the class of $\tilde{I}$. The computation of $D_{(M_F,C)}(\tilde{I})$ gives the position of $\tilde{I}$ with respect to C. If we know the relative position of every class C' and even if only one distance is computed, we can estimate the value of the membership function for each class. Indeed, if the distance between C and C' differs from the distance between C and $\tilde{I}$, the probability that $\tilde{I}$ is a member of C' is low.

More formally, for each class C', the set of inter-class distances $DX_{M_F,C}(C')$ between $I^*_{M_F,C}$ the representative of C according M_F and images from the C' is computed. $D_{(M_F,C)}(I)$ is the distance between the representative of C et $\tilde{I}$. Then, $\tilde{I}$ might be a member of every class C'' where $D_{(M_F,C)}(I)$ is close to $DX_{M_F,C}(C'')$.

This is the main idea of our approach. In classical methods, $\tilde{I}$ is considered as a member of the nearest class in the feature space. Our approach computes

the distance d between $\tilde{I}$ and a point of view. $\tilde{I}$ is then considered as a member of a class whose distance to the point of view is close to d.

The classification module has $Card(MetricSet) \times Card(ClassSet)$ couples (M_F, C). Each of these can be a point of view. Then the training step of our module consists in computing the distances between the document images from the different classes.

The classification of an unknown document consists in selecting a sequence of couples (M_F, C) and exploiting the different classification hypothesis to identify the real class.

5 Learning

The learning phase is performed for each feature F and each corresponding metric M_F. First, a document image from $ImageSet_C$ is chosen for each couple (M_F, C) to represent the C class according M_F. This is performed by computing the intra-class distances $D_{M_F}(I, J), \forall (I, J) \in ImageSet_C \times ImageSet_C$. We deduce $I^*_{M_F,C}$ from the intra-class distances. Different choice might be implemented. The representative can be the gravity center, the image which minimize the greatest intra-class-distance or the image which minimizes the standard deviation of the intra-class distances. This choice may be the subject of future study. Our choice is the image which minimizes the standard deviation.

The second step of the learning phase consists in giving to each class the knowledge of the distance which separates it from other classes according to each metric. To do so, we could evaluate the distances between every document image from a C class and every document image from C' classes. But to reduce the training time, only the distances between the class representative of C and every document image from the other classes are computed. Theses distances allow to build the sets $DX_{M_F,C}(C')$ and $DI_{M_F,C}$ of inter and intra-class distances (cf. notation 3).

Each class has now the knowledge of the distance between its representative and every image from the other classes according every M_F.

Then, a membership function can be defined from the computed learning data. Indeed, according to each feature, each class has the knowledge of the distance with the other classes.

So if a document image is represented by a point in the feature space, and if the projection of this point in the distance space is located in the interval $[\min(DX_{M_F,C}(C')), \max(DX_{M_F,C}(C'))]$, then the image might be a member of the class C'. On the other hand, the more the point is far of this interval, the lower is the value of the membership function.

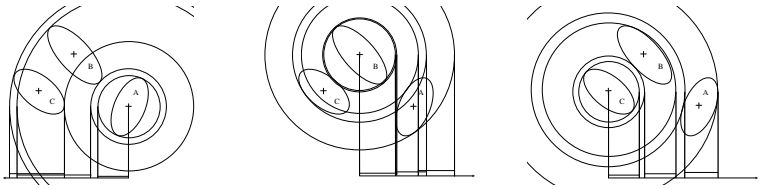
6 Classification

Classification is performed by iteratively filtering the list of candidate classes.

First, for each iteration, a (M_F, C) couple is selected. The module selects the couple for which the ranges of the $DX_{F,C}(C')$ are the most different for every C'

in the candidate classes. The most informative couple is the one which maximizes the distances between intervals and which minimizes their intersection. These intervals in the distance space correspond to hyper-rings in the feature space (see Fig. 1(a), 1(b) and 1(c)). In other words, the couple (M_F, C) which gives the best point of view, if it is the one with the lowest intersection between hyper-rings. Then, an image represented by a point in the distance space, will belong to a low number of intervals.

Figures 1(a), 1(b) and 1(c) presents a problem with 3 classes and 1 feature space. It illustrates the discriminating power of different points of view. Indeed, if the selected point of view is the center of the class A, there is an intersection between the ring including the class B and the ring including the class C. This means that the corresponding intervals in the distance space overlap. The interval overlap is lower if the selected point of view is the center of class B. There is no more overlap if the center of class C is chosen. Then each point of the feature space belongs to one ring at most. This choice can be made in different feature spaces and with different metrics.



(a) (b) (c)

Fig. 1. Choice of the point of view

The distance $D_{(M_F,C)}(\tilde{I})$ between $\tilde{I}$ and the selected point of view $I^*_{M_F,C}$ is then computed with the metric of the selected feature F. This distance computation corresponds to draw a circle in $Space_F$ centered on the representative image of the selected class C. All classes which intersect the circle are kept in the candidate class list because the membership function is equal to 1. Classes whose membership function value is lower than a threshold are removed from the list.

This process is repeated until one class or less remains in the candidate list. The selection of the best point of view correspond to choose the feature space and the metric with the best discriminating power for the candidate classes. The iterative process corresponds to determine a path in a dynamically built decision tree [4]. If no class remains, $\tilde{I}$ is rejected.

The choice of the best point of view does not take into account the eliminated classes. However, their representative are still potential points of view.

```
Candidates ← ClassSet
while Card(Candidates) > 1
 choose best couple(F,C)
```

```
 compute d= D_(F,C) (I)
 remove far classes in Candidates according to DX_F,C
endWhile
if Card(Candidates) > 0
 then C=Candidates[0]
 else reject I
```

7 Experimental Results

The classification module contains 6 features (graphical [2] [5] and structural [1] [6]). For the presented test, the representative of each couple (M_F, C) is chosen as the image which minimizes the intra-class distance standard-deviation.

The tests have been performed on different bases containing images representing differents classes of cheque-deposits, forms and air flight coupons. Experimental results are given in Table 1.

As mentionned in section 2, the classes are defined by the position and the nature of the handwritten fields to be read. The preprocessing (eg. binarization) is an other source of variability. On the other hand, users only provide few images for learning (5 to 10 images per class). Fig. 7 shows examples of images from two different bases.

	#Images	#Classes	#Iterations	Good classification rate	Reject rate	Confusion rate
A	25125	2	1	100%	0%	0%
B	123	8	2 to 4	96%	4%	0%
C	65689	6	2 to 3	92%	4%	4%
D	8770	6	2 to 3	95%	4%	1%
C+D	10000	12	2 to 4	89%	1%	9%

Table 1. Results

The results presented in table 1 have be compared with the ones given by a $k$nn classifier. In a first experiment, tests have been performed with a classifier for each of the 6 features. Several values of k have been tested (1, 3, 5 and 10) but it seems that it has a low influence on the results. The results also showed that some features are well suited for a problem and have a low discriminating power on an other. Indeed, the best configuration gave a classification rate of 96% on a problem, but the same feature gave classification rate lower than 60% on an other base.

In a second experiment, all the features have been grouped in a single feature vector. The results were lower than 40% for every base, the discriminating power of well suited features being disturbed by others. A feature selection should have been performed for each base.

This comparison validate the dynamic feature selection of our approach. Moreover, the $k$nn classifier needs to compute the distance with every point of the training set whereas our approach only compute a distance per iteration.

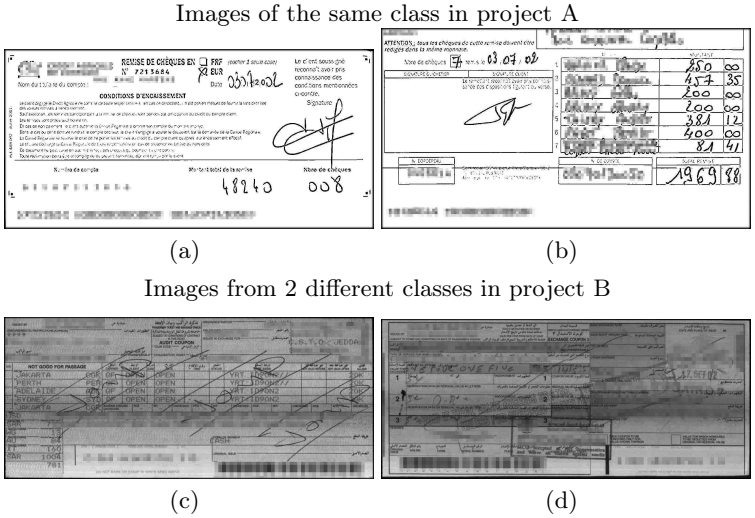


Fig. 2. Exemple of images

8 Conclusion

This article presents an original classification strategy which is applied to document image classification. This strategy is implanted as a module of a complete solution which aims at automatically reading handwritten fields on document images from multiple classes.

The strategy presents some advantages. First, the classification process is a sub-linear function of the number of classes. We do not have to compute the distance with every class representative. This allow to use it in an industrial context. Moreover, the module is easily configurable. By setting the number or the proportion of candidate classes to eliminate, a maximum number of iteration can be set. Lastly, the module architecture allows to add features very easily. In our system, a feature is only defined by its features space and its distance computation. Indeed, it is only required to write a function that places an image in the applied feature space and a distance function returning a real in $[0, 1]$. So for a particular difficult project with specific constraints, we can easily add features dedicated to the discrimination of the not easily separable classes.

At each iteration of the classification process, the best feature vector with the best point of view is dynamically determined.

The main advantage to work in distance spaces, is that it allows to use a large variety of feature. The processed feature vectors can be very different. There is no restriction concerning its size and nature. Numeric, syntactic and structural features can be mixed in the feature set and even in the same feature vector if a distance can be computed between two vectors.

The learning phase can be incremental. If a new class is added, the previous computed distance are kept, only the distances with the new documents have to be evaluated. New features can be added in the same way.

It seems that this strategy can be applied to a large scope of classification problems with supervised learning.

The strategy presented in this paper is a very preliminary work. The first results seem to be promising, but they could be improved by further works on many aspects (choice of the class representative, dynamic determination of the best feature vector for the best point of view).

Another future work would be to improve the used feature set. For example, the better feature vector could be determined to limit the computation complexity. The feature vectors could be combined to build new ones, for example by using genetic algorithms. We can imagine that the built feature vector would use features exploiting different zones of the document image. The genetic algorithms would then select the features associated to the most discriminating zones.

References

1. Héroux, P., Diana, S., Ribert, A., Trupin, E.: Classification methods study for automatic form class identification. In: 14th IAPR International Conference on Pattern Recognition ICPR'98, Brisbane, Australie, International Association on Pattern Recognition (1998) 926–928
2. Clavier, E.: Stratégies de tri : un système de tri des formulaires. PhD thesis, Université de Caen (2000)
3. Ribert, A.: Structuration évolutive de données : Application à la construction de classifieurs distribués. PhD thesis, Université de Rouen (1998)
4. Vannoorenberghe, P., Denoeux, T.: Handling uncertain labels in multiclass problems using belief decision trees. In: IPMU'2002. (2002)
5. Unser, M., Aldroubi, A., Gerfen, C.R.: A multiresolution image registration procedure using spline pyramids. In: Wavelet Applications in Signal and Image Processing. Volume 2034., SPIE (1993) 160–170
6. Azokly, A.S.: Approche uniforme pour la reconnaissance de la structure physique des documents composites baése sur l'analyse des espaces blancs. PhD thesis, Université de Fribourg (1995)