Classification Method Study for Automatic Form Class Identification

Pierre Héroux, Sébastien Diana, Arnaud Ribert, Eric Trupin

Classification Method Study for Automatic Form Class Identification

Pierre Héroux[1], Sébastien Diana[1,2], Arnaud Ribert[1], Eric Trupin[1]
[1]Laboratoire PSI, Université de Rouen, Place E. Blondel, 76821 Mont Saint Aignan Cedex, France
[2]DPCi S.A., 15 rue J-B Colbert BP 6042, 14062 Caen Cedex
{Pierre.Heroux, Sebastien.Diana, Arnaud.Ribert, Eric.Trupin}@univ-rouen.fr

Abstract

In this paper we present three classifiers used in automatic forms class identification. A first category of classifier includes the k-Nearest Neighbours (kNN) and the Multi-Layer Perceptron (MLP) classifiers. A second category corresponds to a new structural classifier based on tree comparison. On one hand, a low level information based on a pyramidal decomposition of the document image is used by the kNN and the MLP classifiers. On the other hand, a high level information represents the form content with a hierarchical structure used by the new structural classifier. Experimental results are presented. Some strategies of classifier co-operation are proposed.

1 Introduction

A form processing system automatically extracts and understands the content of the forms. Such a system is based on the knowledge of location and meaning of areas on the form and consistency links between them which define the *reading model*. Most of form processing systems have to be set up by the reading model to process a given form. Our approach is based on an automatic form class identification to select the reading model (Figure 1).

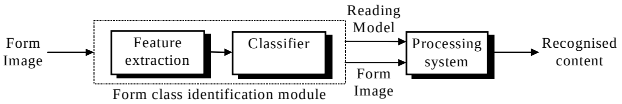

Figure 1 :Form processing system evolution

The identification module extracts a low level information used by classical classifiers, and a structural representation of the form content used by a new classifier.

In section 2, the information extraction module is developed. Classification methods using the low level information is presented in section 3. Section 4 presents a new structural classification method based on tree comparison. Experimental results concerning each classifier are detailed in section 5. Finally, strategies of hybrid approaches are presented in section 6.

2 Information extraction

In this section we present the two kinds of information extracted by the identification module.

2.1 Image pyramidal decomposition

This pyramid construction is based on a recursive cut of the binary image into rectangular regions in which black pixel density is calculated. The pyramid presents several cut levels with a different granularity. The first level corresponds to the black pixel density in the whole image, the second level gives black pixel density in 4 rectangular cuts, the third level cuts the image in 16 parts… A 5 level cut returns a 341 (1+4+16+64+256) feature vector used by the MLP and kNN classifiers.

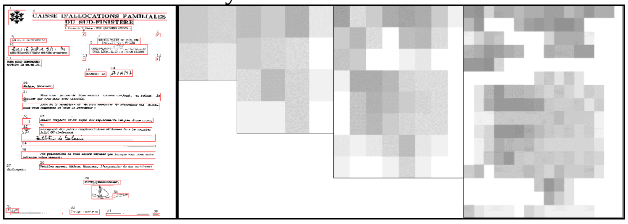

Figure 2 : a form and its 5 level cut

2.2 Hierarchical structure extraction

A high level information describes the organisation of the form content with a hierarchical structure (modelled by a tree) which represents the hierarchical links between each element of the form (Figure 3).

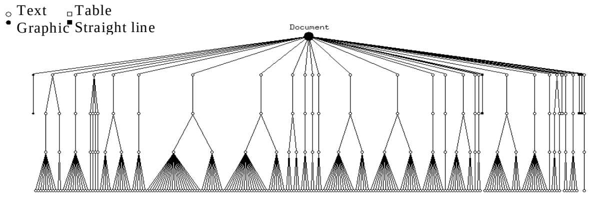

Figure 3 : hierarchical structure

The extraction is organised in five main processes. A segmentation phase is performed on the form image and returns homogenous blocks [1] [2]. Then, the blocks are labelled (text, graphic, line segment, table). The text-blocks are segmented in sub-blocks representing text-lines. The relative location of text-lines in the block allows to perform a logical labelling of the text-blocks (paragraph first line, paragraph line, paragraph last line…). The extracted elements are organised in order to obtain a tree structure in which nodes are labelled (location, dimension, layout and logical attribute). For more detailed information on structure extraction see [1] et [2].

3 Two classic classification methods

The k-Nearest Neighbours [3] and the Multi-Layer Perceptron use as an input the feature vector expressing the pixel density at different resolution. The principle of classifiers are reminded in the following paragraphs

3.1 K-Nearest Neighbours

The kNN classifier is based on the fact that if two elements are close in their representation space, they probably belong to the same class. As mentioned before, each form is represented by a feature vector which represents its co-ordinates in a multi-dimensional representation space. In order to identify the class of a form, the form class of its k nearest points in the representation space is a significant information.

It is commonly admitted that the reliability increases with k. A strict rule is to impose that the k nearest neighbours belong to the same class to take a decision. This rule reduces the number of errors, but it leads to many rejects. It can be smoothed by weighting the voting of each neighbour according to its rank or distance. The main drawback of this classifier is the number of calculations needed to find the k nearest neighbours. Some improvements are proposed in [4].

3.2 Multi Layer Perceptron

The MLP is a layer-organised neural network. Only connections between neurones of two consecutive layers are allowed. The neurone output is calculated as the weighted sum of its inputs to which is applied a non-linear sygmoid function. Synaptic weights are usually determined in a learning phase using back-propagation algorithm [5]. The MLP is known for its decision speed and its good generalisation capacity.

4 Structural classification method

This method performs the form class identification by comparing a tree representing the form (section 2.2) with tree representation of the form classes.

Many tree comparison methods exist. Some of them (Selkow’s algorithm [6], Thomasson and Gonzalez [7]) are based on an isomorphism and label similarity measurement. Others are graph matching algorithms [8] [9] which can be applied to trees (particular cases of graphs).

Finally, we propose an iterative algorithm. Roots of the trees to be compared are first examined. If they are equal, equality is looked for among their sons. Then the sub-trees whose roots are equal are compared, and so on… Nodes are considered as equal if the difference between their label does not exceed a threshold. Finally, this gives the largest tree common to the compared trees.

Section 4.1 presents the form class identification method. The construction of trees representing form classes is detailed in section 4.2. An organisation of model tree database which improve identification in term of computation time is presented in section 4.3.

4.1 Form class identification

To identify the class of a given form, comparisons between the tree representing the form and the model trees representing all classes are performed. Each comparison returns the common tree between the specific and the model tree. Three features are extracted at each comparison : the number of nodes of the common tree and the overlap rates of the common tree in the compared trees. These rates limit the density variation influence between classes and the variability problem in each class. The examination of these features allows to determine the nearest model tree. The features of the selected model tree are submitted to a threshold presented in 4.2. When the features extracted from the tree comparison do not satisfy the threshold, the form is considered as belonging to an unknown class.

4.2 Model trees construction

Each form class is represented by a model tree. It includes the most frequently encountered features in trees representing forms from the considered class.

The forms are hand-filled. Hand-written data are added in predefined areas called « active areas ». Data variations and writers multiplicity involve a low stability on corresponding nodes contrary to passive areas. Model trees are constructed in a learning supervised phase. In a first step, the most frequently encountered nodes in a training set containing trees from the same class are listed. Nodes which are not significant enough (low appearance frequency) are removed from the list. In a second step, model trees are constructed from this list by linking nodes of two consecutive levels with compatible labels.

Once constructed, the features presented in section 4.1 are extracted from comparisons of trees of the training set with the model tree obtained. A statistical study has shown that the features follow a Gaussian law. For each model tree, mean m and standard deviation σ are calculated. The identification process determines the class for which the model tree is the nearest from the input tree. The returned class is kept as the identification result if the features are over m-2σ, else, the form is rejected as belonging to an unknown class. The automatically computed threshold differs for each class and represents the difference of density and stability in and among the classes.

4.3 The hierarchy of models

The form class identification is improved with a hierarchical organisation of the model tree database which reduces computations. The hierarchy is constructed by recursively grouping common features in meta-models. The obtained hierarchy is a binary tree where the non-terminal nodes correspond to meta-models and the terminal nodes correspond to models.

The form class identification is performed by finding a path leading to a terminal node. The path search is performed at each node of the hierarchy by comparing the input tree with the two meta-models to find the nearest one. An algorithm which allows to simultaneously explore the k best paths is used to make the identification more reliable.

The path search involves a lower comparison number than for the whole model database exploration. Moreover, because of its lower number of nodes, comparing a tree with a meta-model is faster than with a model.

5 Experimental results

In this section, the experimental results concerning the discussed classifiers are presented.

5.1 Classical classifiers

The results (Table 1) concern 570 forms from 27 different classes. This set has been cut in equally sized learning and test sets. For each form 2, 3, 4 and 5 level pyramids have been generated. The kNN classifier has been set with k=1 and uses Euclidean distance. The MLP has been set with 1 hidden layer containing 27 neurones.

Pyramid	Recognition (%)		Reject (%)		Error (%)	
Level	kNN	MLP	kNN	MLP	kNN	MLP
2	96.84	94.04	0.00	1.75	3.16	4.21
3	99.65	99.65	0.00	0.00	0.35	0.35
4	100.00	99.65	0.00	0.35	0.00	0.00
5	100.00	100.00	0.00	0.00	0.00	0.00

Table 1 :Classical classifier results

5.2 Structural classifier

The results (Table 2) have been observed on a 1420 form set. 1300 of these belong to 26 learned classes and 120 belong to unlearned classes. Otherwise, the learning phase has been performed with 10, 20, 30 or 40 forms per class. These results show a good capacity in rejecting unknown forms and no error in the identification of forms from known classes with an interesting recognition rate.

Trees in the	Known classes (%)			Unknown classes (%)	
training set	Recogn	Reject	Error	Reject	Error
10	87.31	11.54	1.15	100	0.00
20	94.62	5.38	0.00	100	0.00
30	97.31	2.69	0.00	100	0.00
40	99.23	0.77	0.00	100	0.00

Table 2 :Results of the structural classifier

The number of forms used in the model tree construction seems to be an important parameter (better performances when it grows).

6 Conclusion

In this paper, we have presented three different classifiers used in form class identification. Classical classifiers (k-Nearest Neighbours and Multi-Layer Perceptron) use a low level information on the binary image of the form. A new structural classifier exploits a tree structure representing the form content organisation. This classifier is based on tree comparison.

The three classifiers show good recognition rates. Moreover, these results highlight the pertinence of the used features even if it will be necessary to test their performances on more representative test set.

Finally, our prospects are based on the set up of a classification strategy which consists in combining both classifier types. A first strategy corresponds to use the classical classifiers as pre-classifier to reduce the number of candidate classes. Then, the structural classifier will be applied to find the correct class. A second approach consists in co-operation where both classifiers suggest an ordered list of candidate classes from which correct class is then selected by a vote.

7 Acknowledgements

This study was performed in the ADI (Intelligent Archiving and Documentation) project context in association with DPCi, a company specialised in document scanning, printing and processing.

8 References

[1] S.Diana, E.Trupin, Y.Lecourtier and J.Labiche, *An Assistant to the Modelisation of Forms*, MMSP'97, pp. 163-168, 1997.

[2] S.Diana, E.Trupin, Y.Lecourtier, and J.Labiche, *From Acquisition to Modelisation of a Form Base to Retrieve Information*, ICDAR'97, pp. 762-765, 1997.

[3] T.M.Cover, and P.E.Hart, *Nearest Neighbour Pattern Recognition*, IEEE Trans. on Information Theory, Vol. 13(1), pp. 21-27, 1967.

[4] K.Fukunaga, *Introduction to Statistical Pattern Recognition*, 2^{nd} Edition, Academic Press Inc., Boston, 1990.

[5] D.E.Rumelhart and J.L.McClellard, *Parallel Distributed Proc. :Explorations in the Microstructure of Cognition*, Vol. 1 :Foundations, The MIT Press, 1986.

[6] L.Miclet, *Méthodes structurelles pour la reconnaissance des formes*, Ed. Eyrolles, 1984.

[7] M.G.Thomasson and R.C.Gonzalez, *Syntactic Recognition of Imperfectly Specific Patterns*, IEEE Trans. on Computers, vol. 24(1), pp. 93-96, 1975.

[8] Y.Ishitani, *Model Matching Based on Association Graph for Form Image Understanding*, ICDAR'95, pp. 287-292, 1995.

[9] A.Budin, *On the Problem of Attributed Relational Graph Matching*, Automatika, n°33, pp. 151-157, 1992.