Bayesian networks learning algorithms for online form classification

Emilie Philippot, Yolande Belaïd and Abdel Belaïd
University Nancy 2 LORIA - Campus scientifique - BP 239
54506 Vandoeuvre-lès-Nancy Cedex - France
{*emilie.philippot, yolande.belaid, abdel.belaid*}*@loria.fr*

Abstract

In this paper, a new method is presented for the recognition of online forms filled manually by a digital-type clip. This writing system transmits only the written fields without the pre-printed form. The form recognition consists in retrieving the original form directly from the filled fields without any context, which is a very challenging problem. We propose a method based on Bayesian networks. The networks use the conditional probabilities between fields in order to infer the real form. Two learning algorithms of form structures are employed to test their suitability for the case studied. The tests were conducted on the basis of 3200 forms provided by the Actimage compagny, specialist in interactive writing processes. The first experiments show a recognition rate reaching more than 97%.

1 Introduction

The work reported in this paper addresses the problem of form classification filled out manually using digital pens. This research is undertaken in collaboration with the Actimage company which is our partner specialist in interactive systems. Actimage is looking for a solution concerning the notetaking by using digital pen with clips.The paper used is a conventional and standard paper . The writing option on electronic paper (e-paper) has been removed to improve the flexibility and reduce the cost. Furthermore, this choice have been made for a miximum of generality. In fact, this capture technique is very used for other applications such as correction of plans or taking notes. The digital pen is replaced by a transmission system of electronic ink without sending the pre-printed form. The use of this kind of input mode is important for the company because it accelerates the filling procedure. However, the automatic recognition of the form class becomes complex due to the context loss. Figure 1 shows an example of the problem. At the top left, we see the completed form. In the upper right is the field transmitted to the system. Below, are the different forms which are candidates for the recognition.

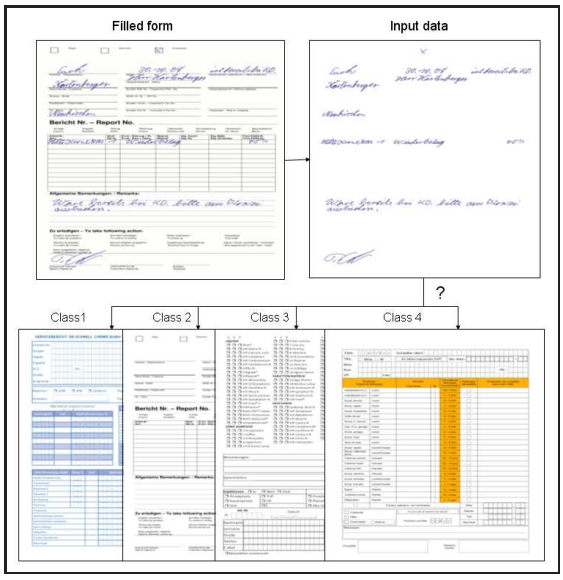

Figure 1. The research problem

The literature shows, for form classification, a lot of research mainly oriented towards off-line forms where information about the context is present in the document image. As an example, we can mention [5] where the document is first segmented into physical blocks using the white spaces, then a probability distribution of the block locations in the image surface is trained. At last, the form recognition is performed using the minimum distance between the distribution of the input form and all the distributions learned. In [2], here also, the authors use the pre-printed form. From this model, they look for the blocks by locating their corners. In case of no matching, they operate on them, some reorientation and scaling corrections. This matching allows them to additionally detect some writing errors.

[7] propose an alternative method of form classifica-

tion also based on the structure extraction, using decision trees. There are trees of local structures and global tree structures representing the aggregate forms at both levels of the hierarchy.

In [6], Neschen offers a system for automatic reading of German bank forms. His approach is based on form segmentation and on the use of a classifier (nearest neighbor classifier) and finally on a correction unit. The interest of such technique is based on the device correction that improves the recognition.

Concerning the online forms, the literature mentions only researches related to word recognition without considering the form classification aspect. It is why we propose a new approach for the online form classification. However, we can inspire ourselves work on segmentation into blocks, matching fields and probability distribution for this classification.

In [8], an approach is proposed for online multi-strokes composite sketchy shape recognition. A classifier using a double-level Bayesian networks is designed to model the intrinsic temporal orders among the strokes effectively, where a sketchy shape is modeled. The drawing-style tree is then adopted to capture the users' accustomed drawing styles and simplify the training and recognition of Bayesian network classifier.

The authors consider here [1] the task of structured document classification. They propose a generative model able to handle both structure and content which is based on Bayesian networks. They then show how to transform this generative model into a discriminant classifier using the method of Fisher kernel. The model is finally extended for dealing with different types of content information (here text and images).

The paper is organised as follows: first, section 2 describes the proposed approach with the different phases concerning the field extraction, the Bayesian network training and form recognition. In section 3, the first results will be presented before we conclude in the last section.

2 The system overview

The approach is partly based on the observation that there are dependencies between fields in a form and between fields and the form. For example, boxes representing "Mrs.", "Mr." and "Miss" in the address area of a form are never checked simultaneously; the presence of a customer identification number implies the absence of customer identify.

Figure 2 shows a dependency example that may exist for the address area. Links and probabilities are used to locate and quantify the dependencies that exist between fields themselves and between the fields and the class. For example, the *city* field depends on the field *zip code*. The table in the lower right corner shows that if the *zip code* is filled, the probability will reach 0.7 meaning that the *city* field will also be filled. Conversely, if the *zip code* field is empty, this means that *city* field will be also empty with a probability reaching 0.8.

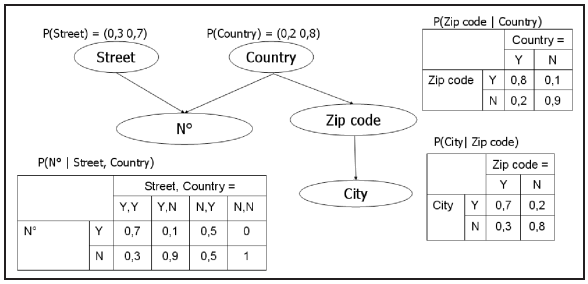

Figure 2. Example of Bayesian network for the address area of a form

Furthermore, the system can have several artefacts that complicate the interpretation task. The data, which is produced by a list of drawn strokes, can be incomplete, ambiguous and can overlap fields. Considering these artefacts, we opted for the use of Bayesian networks because they allow us a qualitative and quantitative dependencies and uncertainty managing. We use a hierarchical approach considering the Bayesian networks by areas of interest, from local blocks (address client, agency information, etc.) until global form. We chose to divide each form into 3 areas of interest corresponding respectively to the heading, the body and to the footer area. These zones cover the three parts usually present in the forms such as the identification area, the filling area and the validation area. This network hierarchy offers some advantages : the reduction of the number of variables trained ; several forms areas can be represtented by the same network ; only affected areas must be re-trained.

Once the Bayesian networks representing the form areas are learned, they are gathered to train a Bayesian network for the classification of the entire form.

Each form class is represented by a model. It is a list of fields where each one is represented by its bounding box, its type (checkbox, string, number, etc.) and the area to which it belongs. We use an XML format to describe the model. This model will serve as a basis for the field extraction.

2.1 Field extraction

Fields are written by hand using the digit pen. They are represented by a list of strokes composed each one by a list of 2D points in the surface of the form. The

system proceeds to the field extraction in two phases. First, the stroke point coordinates are compared to those of the field bounding boxes in the model form. Then, if a majority of them (fixed experimentally to 85%) belongs to the bounding box of a model field, we consider that the field is filled. Once all the strokes are treated, if 20% of them have not been matched, the tested model is excluded.

Figure 3 shows an example of different possibilities of stroke interpretation for the single word "Jean Claude" depending on the context. In case (1), the text is associated with two checkboxes and a text field. In case (2), the strokes are associated with two different text fields. Case (3), which corresponds to the reality, shows that the stroke exceeds the scope which is too small. The challenge will therefore lie to find form initially completed even if the stropkes do not correspond fully to the fields of the latter.

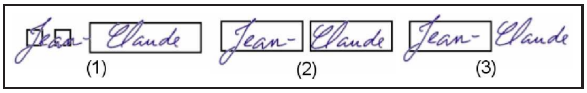

Figure 3. Interpretation example of strokes according to the context

2.2 Bayesian network Training

The training takes place in two stages. First, the main areas corresponding to the most important substructures of the form are identified manually and presented to the system to initiate the Bayesian networks accordingly. Then, the training is continued for the entire form.

For area training, a fully connected graph is constructed from the list of its fields. This constitutes as a basis for the training of the Bayesian network structure of this area. The fields represent nodes in the graph. Within each node, there is a probability distribution qualifying in a quantitative manner the interaction between nodes. The arcs represent the dependency between fields. We use two different algorithms for structure training, PC and MWST [4] in order to test which of the two is best suited to our problem. The PC algorithm consists in testing the conditional independence between variables to generate a graph representing them. The MWST algorithm is an algorithm for finding tree of maximum weight. To each arc is assigned a weight. Then, we seek the tree of maximum weight.

Once the Bayesian networks is trained for all the areas of the form with particular distribution probabilities, the training is enlarged to the entire form by gathering the different Bayesian networks. We apply the PC algorithm and MWST in order to determine the structure of the global network. The global network is the network that summarizes the relationships between areas representing all classes of forms. Figure 4 shows an example of a global network obtained from the classification of two form classes. It is observed that certain areas may refer to two separate forms.

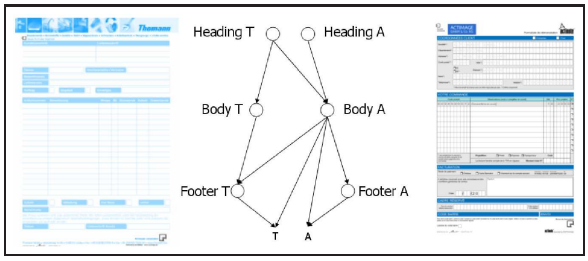

Figure 4. Example of global Bayesian network allowing classification for forms of classes A and T.

2.3 Recognition

Recognition takes place in several stages. First, as in the training field, fields are first extracted by matching the strokes with the form field models. Then, for each area, a belonging probability is performed, i.e the probability that the extracted fields belong to a given area using the networks for defined areas. Finally, the different probabilities obtained are used to constitute the form's class with the global network.

3 Experiments

We experimented a database provided by the Actimage company. Currently, we have limited our experiments to 4 classes in order to restrict the run time and validate our approach. Subsequently, this number will be increased for more experiments the closest possible to the industrial operating environment. These four classes of forms are presented in Figure 1 and contain 800 samples per class. 600 forms are used for training and 200 for recognition tests using a cross validation method. Each network is trained on 4 different learning bases extracted from 3200 forms in the initial sample. The tests were performed in Matlab using the BNT toolbox [3].

From a global view the results presented in Tables 1 and 2 are encouraging. Using the algorithm MWST we get a global recognition rate of 97.89 %. The PC algorithm gives a recognition rate of overall 90.76 %.

Class	Algo	Heading	Body	Footer	Global
1	PC	99,38	98,5	99,25	96,88
	MWST	61,06	91,53	77,63	98,83
2	PC	99,7	66,88	98,5	91,8
	MWST	58,02	87,79	41,96	98,63
3	PC	0,13	89,13	2,25	98,75
	MWST	57,93	92,47	53,81	96,25
4	PC	99,02	99,21	0.13	75,63
	MWST	61,12	86,67	26,42	97,88
Avg	PC	74,56	96,63	50,03	90,76
	MWST	59,53	89,61	49,95	97,89

Table 1. Recall (in %)

Class	Algo	Heading	Body	Footer	Global
1	PC	56,68	98,5	37,62	80,2
	MWST	74,5	89,75	50,38	98,62
2	PC	83,86	66,88	74,7	99,6
	MWST	74,88	87,25	50,12	98,25
3	PC	25	89,23	0,5	95,98
	MWST	74,5	90,75	50,5	95,96
4	PC	97,66	98,88	0,13	98,77
	MWST	74,5	85,75	49,12	97,53
Avg	PC	65,8	88,37	28,23	93,9
	MWST	74,6	88,37	50,03	97,59

Table 2. Precision (in %)

Regarding the results on form regions, the algorithm MWST gives more homogeneous results than the PC algorithm. Indeed, the recognition rate and accuracy are constant whatever the basis of learning. However, the PC algorithm can achieve very good results in certain areas. This can be helpful in cases where a form is completed in several stages with intermediate treatment between each of these steps.

For the PC algorithm, we note that the accuracy rate of class 1 which is only 80.2% for the global recognition, is due to the complexity of its structure. Indeed, its fields are short, numerous and very close. The extraction step of the fields is strongly biased by this peculiarity as a text field from another class will cover several fields of class 1. For example, we note that the recall rates of the class footer 4 is only 0.13%. Similarly we observe that the precision rate of the footer of the class 1 is only 37.62%. This is explained by the overlapping fields in two classes. The fields of class 4 *footer* are completely subsumed by the fields of class 1 *footer*. The matching is biased and the recognition of the form area are distorted. Nevertheless, the recognition rates of classes 1 and 4 are good, since the global network accepts the possibility that a class is defined by an area outside its original model. This problem is significantly mitigated by using the algorithm MWST.

4 Conclusion

We have developed and tested a first approach for the classification of online and unconstrained forms by using two levels of Bayesian networks. The approach exploits the conditional probabilities between area fields and strokes in the fields to find the more close form model. Early results are encouraging and pave the way for many opportunities. In the future, it would be interesting to validate the robustness of our system with a larger number of classes. In view of the different results obtained from two structure learning algorithms, it might be interesting to see the contribution of other algorithms both at local and global forms. Then, we plan to test the limits of the system about the direction sheet, and to modify the stroke matching approach by proceeding to a segmentation stage to reduce the matching errors.

Finally, the use of Bayesian networks on forms could be a way to explore new strategies for filling them and thus allows us the modification of the layout and editing content of forms to adapt them to the writers.

Acknowledgment

This work is conducted under a CIFRE agreement. We would like to thank the Actimage company which collaborated in this work and has provided the necessary database for training.

References

[1] L. Denoyer and P. Gallinari. Bayesian network model for semi-structured document classification. *Information Processing and Management*, 40(5):807–827, 2004.

[2] D. Doermann and A. Rosenfeld. The processing of form documents. *ICDAR*, 2:497–501, 1993.

[3] K. Murphy. The bayesnet toolbox for matlab. In *Computing Science and Statistics: Proceedings of Interface*, volume 33, 2001.

[4] R. E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall; illustrated edition, 2003.

[5] S. Ramdane, B. Taconet, A. Zahour, and S. Kebairi. A statistical method for an automatic detection of form types. *DAS*, pages 84–98, 1999.

[6] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, 2000.

[7] T. Watanabe, Q. Luo, and N. Sugie. Layout recognition of multi-kinds of table-form documents. *IEEE Transactions on PAMI*, 17(4):432–445, 1995.

[8] L. Z. Z. Sun and B. Zhang. Online composite sketchy shape recognition based on bayesian networks. *LNCS*, 4222/2006:506–515, 2006.