MVA'90 IAPR Workshop on Machine Vision Applications Nov. 28-30,1990, Tokyo

'SYMBOLIC SUBTRACTION' OF FIXED FORMATTED GRAPHICS AND TEXT FROM FILLED IN FORMS

Gerd Maderlechner
Siemens AG

D-8000 Munich 83
F.R.G.

ABSTRACT

The proposed method has the aim to separate the filled in (typed or handwritten) information in a form, from the fixed preprinted textual and graphical information of the blank form. It is an extension of document analysis methods based on connected component (CC) analysis. After text/ graphics segmentation not only the text information is analysed but also the graphics. The CC's classified as graphics are skeletonized maintaining a correspondence between the contour and the skeleton lines. In this paper we use a model of the blank form, describing the layout and logic structure of the text and graphics components. By recognition of the graphical layout the class of the form is recognized. The filled in text is discriminated from the preprinted text and recognized even if it is overprinted on the graphics layout of the form.

INTRODUCTION

Today the automatic reading of fixed format forms is performed by fast reading machines, e.g. in banking applications, using specially prepared forms with preprinted graphics and text in a special color, which is physically segmented by special scanners using a color filter. The problem of separating text overprinted on graphics is avoided.

We propose a new segmentation method that will allow the use of ordinary black and white forms and standard greyscale scanners without color filters. This method performs a segmentation of the preprinted graphics and text of the blank form by intelligent document analysis.

Simple subtraction of the raster image of the preprinted blank form (BF) from the raster image of a filled-in form (FF) as pixelwise image subtraction does not work for the following reasons: 1. Inexact positioning of the form sheet on the scanner causes translation and rotation errors. 2. Different saturation of printing and variations in paper quality cause different varying line thicknesses. 3. Different noise and threshold levels of blank form and filled form. 4. Distortions by scanners and copying.

Even determination and compensation of the mentioned distortions, e.g. of rotation and translation, does not produce a sufficient clean difference image for text recognition.

In analogy to this subtraction we propose the name 'symbolic subtraction' for our new approach in this paper: the 'symbolic difference' Sym(FF)-Sym(BF) means the recognition of the filled-in text using the models of FF and BF.

SYMBOLIC DESCRIPTION

The symbolic description of the *textual components* of the form follows the document analysis method as described in [KLM88]. After CC analysis and text/graphics classification [S87] an object oriented knowledge based analysis yields a hierarchical description of the layout and logic structure of the textual part of the document. From the lowest level to the highest level there are the CC's, building characters, which are grouped to words, lines and text blocks, represented by rectangles with known position and size.

In this paper the *graphics components* of the form are further analysed. The structure of the graphics of the blank forms is characterized by horizontal and vertical lines. By a thinning method solely based on the contours of the graphical CC's [E87] we get the characteristic skeletonal lines. This method preserves the neighborhood relation between the skeleton branches and contour segments. The final graphics representation is a graph structure with nodes as the branching points of the skeleton and arcs as the skeleton lines.

This procedure is applied to the blank form to get the model description and to the filled in form for recognition.

MODELING OF FORMS

The models for the blank and filled forms (BF, FF) are defined in close analogy to the Office Document Architecture (ODA) by the layout structure, logic structure and contentof the document.

The content of FF with its logical meaning (e.g. account number, customer etc.) is the relevant information. The layout and logic structure of BF is fixed for each class of forms, and we use them only for identification of the

class of FF.

The BF is modelled by a text frame, containing text blocks described by position and size, and a graphics frame containing mainly horizontal and vertical lines.

The model of the filled form (FF) is described by an additional transparent text frame with text blocks in geometric and logical relation to the BF. The position and size of each text block has a tolerance of about a character height.

The reference model for recognition of the form class is based on the graphics lines and their relation: number, length and relative distance of horizontal and vertical lines, corners and vertices. The recognition is based on size, translation and rotation invariant pattern recognition [MKH83].

The text areas of the BF are described in relation to the graphics. After graphics recognition the position and rotation of the document (FF) relative to the model of BF is known and is compensated.

TEXT/GRAPHICS SEGMENTATION

Traditional text/graphics segmentations of documents [see e.g. S87, CW90] do not use additional knowledge of the preprinted blank form. Connected component based segmentation cannot separate overprinted text from graphics [S87]. Sophisticated methods for segmenting text merged with graphics are described in [KIA84] using neighborhood line density (NLD), in [BK90] using textline context, and in [KI90] using apriori knowledge of the graphics lines. But these approaches do not discriminate between the relevant filled in text and the preprinted text of the blank form. Using the described symbolic description and modelling of the form, a flexible form recognition is achieved, as descibed in the following section.

MODEL GENERATION AND RECOGNITION

The model of the blank form BF is generated by the following procedure:

Starting from the binary image we get a first symbolic description by the black connected components. A preliminary text/graphics segmentation is performed by CC-feature analysis (Fig. 2 and 4 a,b).

The text components are grouped to the mentioned hierarchical layout structure of text blocks. The logical meaning is labelled interactively.

The graphics components are skeletonized from their contours (Fig. 5b) without using the image matrix [E87]. By polygonal approximation and smoothing [MKH83] we yield a symbolic description by graphical primitives, in our case mainly horizontal and vertical lines, and their interrelationship, characterizing the model of BF.

The recognition process of FF is divided into three parts:

The first part of the recognition process is the same as in the model generation. After the preliminary text/graphics segmentation based on CC´s (Fig. 1 and 3 a,b) the graphics components are further analysed, containing also the overprinted text (Fig. 3 b).

The second part is the further analysis of the graphics components by skeletonization and polygonal approximation, resulting in a skeleton graph (Fig. 5a).

The third part is a graph matching between the skeleton graph of the filled form and the reference graph of the blank form. This process is tolerant against rotation, translation and scaling within limited tolerance of about the size of the smalles distance between two parallel lines. The confidence measure for the recognition of a form class is determined by the ratio of found lines to possible lines in the matching.

If a BF class has been recognized a *refined text/graphics segmentation* is started, using the knowledge of the BF model for graphics and text.

First the graphics is analysed further to restore overprinted text. The comparison of the graphics lines of FF with those of BF identifies additional nodes in the skeleton graph of FF (Fig. 5a). Starting from these nodes the touching characters can be restored, because the contour base thinning [E87] maintaines the relation of the skeleton with its left and right hand contours. This allows the elimination of the graphics parts up to the contourline of the character parts.

Second the geometric transformation of FF against BF can be compensated resulting in an exact positioning of the text blocks relative to the graphics. Now the filled in text can be located and recognized in the layout blocks, defined by the recognized model class.

RESULTS

The results of the described method of symbolic subtraction´ are shown in the Figures 1 to 6 for a complex filled form with a lot of overprinting and broken characters. The result of Fig. 6 shows the contours of the restored characters, without the broken parts. But they are not lost, because the expected filled in text block has a sufficient position and size tolerance. The corresponding logic structure, in this example the account number, has not been shown in these figures.

CONCLUSION

The presented method of 'symbolic subtraction' restores text information which is overprinted on graphics, and discriminates the relevant text from the preprinted text of the form. In addition the class of the form is recognized, allowing the processing (scanning and interpretation) of unsorted stacks of different form sheets.

The models of the different forms can be generated conveniently by scanning of a blank form sheet, automatic layout generation for graphics and text areas, and input of logic information of the different fill in areas by the user. The recognition is robust against translation and rotation of the form which occur by placing them on the scanning area of flat bed scanners.

REFERENCES

[BK90] S. Bow and R. Kasturi, *A Graphics Recognition System for Interpretation of Line Drawings,* in: R. Kasturi and M.M. Trivedi (ed.), Image Analysis Applications, Marcel Dekker, New York, 1990, 37 - 72

[CW90] R.G. Casey and K.Y. Wong, *Document Analysis Systems and Techniques,* in: R. Kasturi and M.M. Trivedi (ed.), Image Analysis Applications, Marcel Dekker, New York, 1990, 1 - 36

[E87] B. Eichhorn, *Skelettierung mit Randinformation,* Diplomarbeit (in German), Technical University Munich, 1987

[KI90] H. Kato and S. Inokuchi, *The Recognition System for Printing Piano Music Using Musical Knowledge and Constraints,* IAPR Workshop on Syntactic and Structural Pattern Recognition, Murray Hill, 1990, 231 - 248

[KLM88] J. Kreich, A. Luhn and G. Maderlechner, *Knowledge-Based Interpretation of Scanned Business Letters,* IAPR Workshop on Computer Vision, Tokyo, 1988, 417 - 420

[KIA84] K. Kubota, O. Iwaki and H. Arakawa, *Document Understanding System,* Proc. 7th ICPR, Montreal, 1984, 612 - 614

[MKH83] G. Maderlechner, P. Kuner and E. Hundt, *Mustererkennung und Musterbeschreibung von Liniengrafik in Zeichnungen,* Proc. 5. DAGM Symposium (in German), Karlsruhe, 1983, 155 - 160

[S87] W. Scherl, *Bildanalyse allgemeiner Objekte,* Informatik Fachberichte 131 (in German), Springer Verlag, Berlin, Heidelberg, 1987

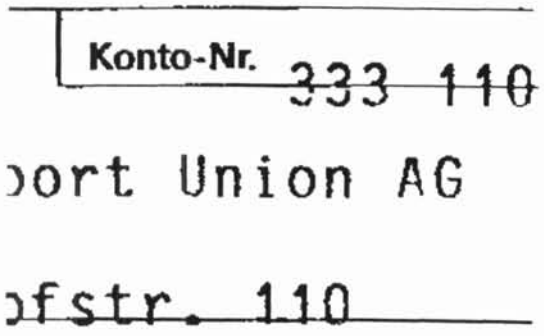

Fig. 1: Detail of an example of filled in form (FF) with overprinting of text on graphics.

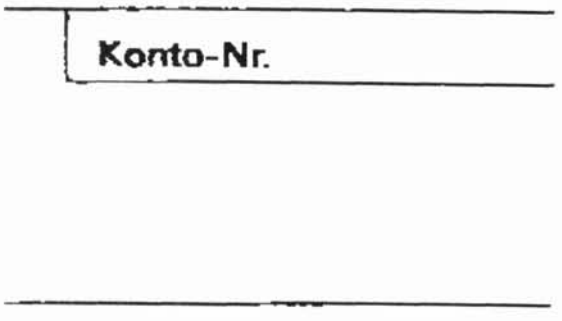

Fig. 2: Detail of blank form (BF) corresponding to Fig. 1 with preprinted text and graphics.

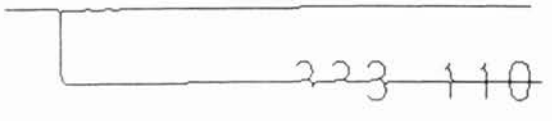

Fig. 5a: Graph of skeleton of graphics components of FF

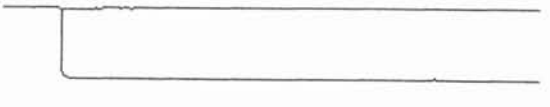

Fig. 5b: Graph of skeleton of graphics components of BF

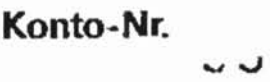

Fig. 3a: Text components of FF after text/graphics segmentation.

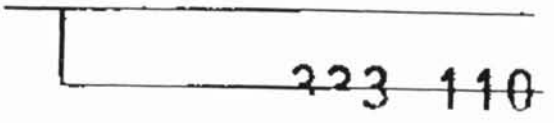

Fig. 3b: Graphics components of FF after text/graphics segmentation.

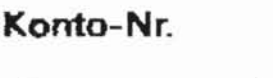

Fig. 4a: Text components of BF after text/graphics segmentation.

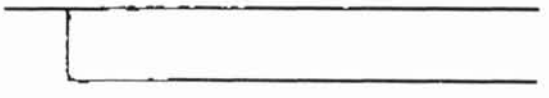

Fig. 4b: Graphics components of BF after text/graphics segmentation.

Fig. 6: Result of 'symbolic subtraction' of the skeleton graph of BF from FF after restoration of the corresponding contours. This information is added to the text components (Fig. 3a) for further text recognition.