Characterizing and Distinguishing Text in Bank Cheque Images

JOSÉ EDUARDO BASTOS DOS SANTOS[1,2], BERNARD DUBUISSON[1] AND FLÁVIO BORTOLOZZI[2]
[1]HEUDIASYC - Université de Technologie de Compiègne(UTC)
BP 20529 - 60205 COMPIEGNE cedex FRANCE
Tel. 33 3 44 23 47 93 - Fax. 33 3 44 23 44 77
{jose-eduardo.santos,bernard.dubuisson}@hds.utc.fr

[2]LUCI[2]A - Pontifícia Universidade Católica do Paraná (PUCPR)
Rua Imaculada Conceição, 1155
80.215-901 CURITIBA – BRASIL
Tel. 55 41 330-1543 – Fax. 55 41 330-1392
{jesantos,fborto}@ppgia.pucpr.br

Abstract: the most common goal of automatic bank cheque treatment systems is the recognition of handwritten information. However, in order to do this, it is necessary to use a reliable and efficient process able to identify and to extract the information, which can then be submitted to a further recognition phase. In this paper we present a process for identifying and distinguishing between handwritten information and machine printed text based on a set of local features. This process is based on the characterization of textual elements via properties derived from their content and their shape. The main advantage of this process compared with other similar approaches is that no a priori information of the treated document is used, thus making it more generic and effective.

1. Introduction

A great majority of research related to extraction of handwriting from bank cheque images is based on aspects linked to document layout. The most commonly employed approaches used are base lines which point to locations on the image where handwritten information can be found, such as the literal amount, or in other cases the use of special symbols, such as currency indicators which point to the region on the image where the courtesy amount can be found. Even though they are simple, contextual approaches have the drawback of being strongly document dependent. Even when applying them to similar documents, some adjustments prove to be necessary. Good examples of this are some cheque models where the base lines are not "solid" but consist of a line formed with reduced size text. Another type of line encountered in some cheques consists of a series of spaced dots. For some systems of handwritten text detection using base lines, these variations can constitute a considerable drawback.

There is another way of detecting handwriting without having to fix environmental aspects. Since humans can distinguish between handwritten text and machine printed text, by simply observing the textual elements, it is possible to design a process based on this same principle.

Human writing can conserve sufficient characteristics to allow its distinction as handwritten text even when it appears on very different types of documents. The ink regularity distribution and the variability of the stroke are basic assumptions on which such differentiation is based. In this way, if it is possible to find a model able to represent these two typographical aspects, it is possible to distinguish handwritten text in a document, in the same way that humans do.

This paper presents some experiments focusing on composing a set of features enabling an efficient distinction between handwritten and machine printed text. Once such features have been observed locally, they can be used for different kinds of documents and situations, thus making the process generic and efficient. As a proof of its abstract nature we have used a database composed of images from various banks and different countries. As shown in the following sections, the results achieved confirm the promise of the methodology employed.

The next section presents a brief review of previous research in this domain. In the third section we introduce the type of features observed in our images and how they are extracted. The fourth section is devoted to the classification phase where we present some of the results obtained. Finally we

finish with a discussion concerning the current stage of this research and future perspectives.

2. Bank Cheque Images Segmentation

Identifying handwritten text on documents is a challenging research area that has occupied ample space in scientific publications especially in the past few years. If the document in question is a bank cheque, a simple way of extracting handwritten elements is to proceed by the subtraction between a filled copy and a blank model[3]. Even though all the printed text, base lines, logo and background are common elements to both images, one has to consider other aspects when implementing a solution based on this methodology. The volume of models that need to be stocked and the spatial distortions produced by digitalizing the images, are questions that cannot be neglected.

When treating Brazilian bank cheques one particular aspect has to be taken into account : the graphical security elements are printed in random positions in each new copy, in order to make eventual falsifications more difficult. Even though these security elements are present in a blank model as well as a filled copy, they are located in different positions in the two images and as consequence of this misalignment, undesired noise appears.

An option for this approach is the successive elimination of some elements of the image rendering handwriting identification difficult. In this process, the background may be the first element to be eliminated. In this way, various solutions have already been proposed, such as adaptations of threshold algorithms [22], morphological based processes[21] or multiscale approaches[8]. Even though in some cases the results can appear to be interesting, degradation of the remaining information is common.

Another element related to elimination which is considered important in such a methodology is the use of base lines. Useful when filling in a document, these represent a significant problem when trying to extract the same handwriting for which they served as a guide. The main impediment to eliminating base lines concerns the points where text and line merge into one single object. Even though the lines can be successfully extracted it is necessary to reconstruct textual elements at the points where they are overlapping, since without doing this, the information represented in the text can suffer serious modifications, making their recognition very difficult or even impossible . Papers such as [11] consider not only base line extraction but also handwritten text reconstruction which is carried out via mathematical morphology.

If in some cases base lines pose problems, in other situations they can be considered as the basis of possible solutions. As pointed out earlier, the base line guides the user when filling in the cheque, hence it is normal to search handwritten text around these lines. Nevertheless some of the principal pieces of research using this approach don't consider models where "non-solid" lines are used.

However, using contextual information as base lines, is not the final option to be considered when locating handwriting text. In fact, to identify handwritten text directly on the image would be a more efficient option for solving the problem, since it exists separately from the other elements on the image.

Djeziri[7] proposed a measurement able to identify all the elements in a document image corresponding to a given filiformity pattern. Handwritten text fits the established filiform criterion perfectly, since the stroke thickness normally doesn't exceed its length, thus conferring on it a "line" or a filiform aspect. However, handwriting is not the only type of text that fits this pattern and as a result machine printed textual elements are also removed in the extraction process.

When aiming to characterize handwritten text, the research of Hobby[2], Cinque[10] and Clark[6] is very relevant here. Hobby's work is based on strokes obtained from image skeletonization in order to distinguish between different kinds of text on the image. The stroke's shape is analyzed and a post processing based on the document's layout generates the final result. The case studies cited concern commercial letters where the main objective is to separate signatures from the rest of the text.

In the case of Cinque's and Clark's work, even though they are not concerned with handwritten text, they employ sets of features for characterizing printed text in order to distinguish them from the rest of the elements in the same location. In fact, as handwritten text can retain some characteristics independently of the document over which it is inserted, the same can be considered as a valid assumption for machine printed text which, except for some rare fonts which are not usually used on bank cheques, conserves sufficient characteristics to enable its discrimination.

In the next section we will see what these characteristics are based on and how they can be used.

3. Selecting Features

A simple way to distinguish printed text from handwritten elements is via an overall survey of these

elements. Printed text has a more regular and homogeneous aspect than handwriting in relation to alignment, spacing and the size of characters. This observation was already the basis for some research conducted in this field, which use bounding boxes over characters in order to observe their alignment and size ratios[23]. However, our objective here is not to observe elements in their global form, but to examine textual elements in small portions, since even they contain sufficient information to allow such a discrimination. As noted previously, ink distribution in the printing process and handwritten text regularity are key factors in distinguishing textual elements. These two aspects represent two fundamental elements when representing text: content and shape. The fact that the content of printed characters is more regular, despite the precision of automatic printing processes, serves as a highly discriminant factor with regard to handwritten text since the ink distribution cannot be controlled during the process of impression, thus causing its content to be more heterogeneous. On the other hand, the shape even when analyzed locally can provide relevant information for discriminating between the two kinds of text.

Let us look at some features linked to content and shape that are used in our classification process.

3.1 Content Related Features

When observing documents where it is common to see darker text on a lighter background, there are two basic aspects to text elements that one should take into account when trying to understand their composition : content and form. As mentioned in the previous section, we assume that a textual element preserves its principal characteristics in many different kinds of documents, under normal situations (original version). In this way, content might provide information about the behavior of gray levels in the textual object. On the other hand, form, in the sense that it is employed here, does not point to an overall relationship of bounding boxes for example, but to a local feature that intends to show the form behavior between different textual elements. In an attempt to construct a compact and efficient group of features, we observed those appearing in geometrical, statistical and shape descriptor fields. However there are many features that are not very useful for our purposes since we are considering machine printed elements composed of a great variety of fonts and sizes.

The use of statistical moments is related to observing how the gray levels behave in textual objects. We note that machine printed elements are more homogeneous in relation to the variability of their gray levels. Normally a single character is not defined as having different colors or gray tones. The aim is always to provide the best possible visual quality so as to make reading easier

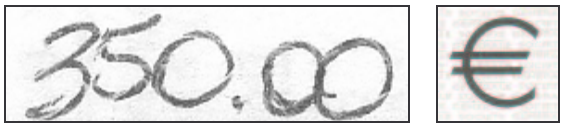


Figure 1 – The two images display differences in content between handwritten and printed text.

However when writing with a pen, we have no control over the ink distribution, thus causing the handwriting to have a greater variability of gray tones. This type of assumption can be used as discriminative information when differentiating between these two kinds of texts. In this case, statistical moments may be used to obtain information from textual patterns.

Range is another useful feature and serves to show the homogeneity of gray levels. We used this in order to verify that the variability interval of gray levels is different for each test category.

Based on the result obtained by the range, we studied a way of amplifying its use regarding not only the different gray levels of the object but also the number of constituent pixels. We call this the *stroke density*, which we define as the ratio between mass and volume, where mass is represented by the number of different gray levels constituting the object and volume is defined by the number of pixels involved. This is a useful feature since it searches for content and also gives an idea of form. Finally, still in relation to content we observed the entropy of the object's gray levels.

Other statistical features can also be used, such as kurtosis, which gives an idea of the degree of distribution of the gray levels in some textual objects.

If one is more interested in searching for machine printed elements which do not change too much in regard to their scale or font style, other kinds of useful moments, such as profile moments or geometrical moments can be employed.

3.2 Shape Related Features

As we set out in the first section, form is another important element of textual discrimination. The main aspect concerning the use of form is not related to a regular pattern for machine printed text from a global point of view, but we assume that even if compared in small parts, machine printed and handwritten text present a sufficient number of differences in regard to their form.

Area and perimeter are simple instances of shape descriptors and are useful when one only intends to search for regular machine printed patterns. Meanwhile, there are other useful shape descriptors which can provide adequate information for our purposes. Observing textual patterns partitioned by a fixed window allows us to note from the elements comprising these windows, some measurements of the smallest ellipse that embodies them, thus acting as a kind of "elliptical bounding box". Eccentricity is one of these potential measurements, and one way to determine this is to calculate the ratio between the lengths of their axes. The value is situated between 0 and 1, where the nearer the eccentricity is to 0, the more the element looks like a line. Another way to measure the elongation of the object is to determine its rectangularity. Rectangularity can be measured as the ratio between the area of the object and the area of its bounding box. The nearer the result value is to 1, the more rectangular is the object.

The linear aspect of an element combined with its shape proportionality favors the distinction between handwriting and machine printed patterns. One way to consider an element's linearity is via eccentricity and an option for observing element shape proportionality involves a feature called solidity. Solidity is bonded with the convex area of the object and is calculated as the ratio between the object's area and the area of its convex hull.

As we mentioned earlier, according to the way in which one analyses textual patterns, there are other features, such as the Euler number or other topological based shape descriptors that can be applied.

In the next section we present a more accurate view of how some of the features above provide discriminating factors in handwriting identification.

4. The Classification Process

In order to extract the features from the image it is necessary to divide it adequately. As the features are observed locally, a square frame of 11x11 pixels supplies appropriate information for our objective. Even so, beforehand, the cheque image is labeled according to its connectivity thus giving an initial idea of the isolation of each object on the image. The labeling is produced through a morphological closing operation with a cross structuring element according to the equation:

$$\Lambda_{B_c}(f)(x) = \begin{cases} \min\{y_1 + Hy_2 : y \in \gamma_{B_c,\{x\}}(f)\} & \text{if } \gamma_{B_c,\{x\}}(f) \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

Equation 1 – Image Labeling

Once the image is labeled the elements are shared via frames of size 11 always retaining only one object per frame. The existence of a single element is important in order to observe shape related features. Displacing the frame through the image and defining the labeled portion of the image to be observed is carried out in the following way : from left to right and from top to bottom, find the first labeled pixel; put this pixel in the center of the frame and isolate the object concerned. After this we define the stroke over which the features will be extracted from the original image. In other words, we use the labeled image only to guide the displacement and isolation process and once we have isolated a stroke we look at their values on the original image in order to calculate the related features.

After extracting the features from this stroke, we take the next labeled pixel not yet having being explored and so on, until all the labeled image has been explored. This sharing phase will generate a data set that will be used as an input to a classifier bounded by distinguishing between two possible classes: handwritten and machine printed elements. After this the data are reduced and centered on the mean.

The classifier employed is a multilayer perceptron trained with a set of 6772 samples extracted from 70 different images. After training, the process was tested with a new set of 5035 samples. At this stage an error rate of 12% was observed and a total of 12% of the samples was considered as ambiguous. For these samples a post processing was necessary in order to determine with more accuracy the class of each of these samples. The incorrectly classified set of samples was composed both by handwritten and machine printed samples. The same balance was not verified on the ambiguous samples where a greater amount of handwritten samples compose this group.

To determine the class of the samples initially classified as ambiguous, an estimation of the probability of a given sample x_i as being part of the ω_1 or ω_2 classes was made ,considering its neighborhood.

The post processing made it possible to reduce the ambiguity rate and to obtain a better quality image as shown in figure 2.

The features employed were selected from a major group and analyzed according to their ability to represent shape and content. The selection was made via a discriminant analysis of this group bounded by selecting the most representative examples.

In relation to possible problems encountered after text extraction presented in [11], we observed that the main problems are related to limits imposed by contextual and extractive approaches, such as:

- Failure to exclude the "$" symbol or overlapping of handwriting;
- Broken strokes or noises caused by the elimination of baselines or bounding rectangular boxes;
- Broken or touched strokes due to a binarization phase;
- Failure to extract the courtesy amount due to absent or broken baselines, or no clear "$" symbol encountered;
- Failure to extract the cents portion positioned on a different line to the dollar portion.

Since in the direct text search we are not concerned either by the "$" symbol, broken baselines or problems caused by binarization phases, we can conclude that this generic methodology gives the text extraction task more freedom.

5. Concluding Remarks

Even though much effort has been spent over the past years in the field of document image segmentation, handwriting identification still remains a challenging task, in particular when distinguishing handwritten from printed text.

Though many different approaches to solving this problem are covered in the literature in this domain, many of them are strongly dependent on document structure. This dependence is an obstacle for applying these solutions to different situations and different documents. Meanwhile, the main elements of a document, for both printed and handwritten text, can retain a sufficient number of characteristics for their identification independently of the document.

Based on this assumption, this paper described a characterization and distinguishing process of textual elements via locally observed features. In this work, textual elements are shared in small portions generating objects over which an adequate set of features are extracted and classified at a later stage.

Being able to observe such features locally and in small portions makes the process independent of the text position as well as dispensing withimage orientation corrections.

The process generality was tested over a database composed of cheque images from different Brazilian banks. Some models of French, Canadian and US cheques were also tested and the results confirm the process quality.

From a general point of view the process is based on two fundamental aspects of graphical composition of a textual element: its content and its shape.

Image 1 clearly shows the composition irregularity of a handwritten text where the user has employed a common pen. The image on the right-hand side displays the homogeneity of gray levels composing the printed character. The relatedness of content and handwritten text elements is represented in this research through features such as mean, standard deviation, stroke density and gray levels entropy. Another important point that needs to be considered when representing these elements is shape, which was observed using features such as eccentricity and solidity.

The way in which these features were observed, i.e., an image labeled according to the connectivity of its elements, leads to a more balanced partitioned image, resulting in a reduction in the error rate with respect to some previous experiments where the image was uniformly shared.

Though the results obtained are promising it is still necessary to carry out a further detailed study in order to obtain an ideal group of features. Another important point concerns implementing a classification process which would be more effective, preferentially allowing the procedure to be carried out exclusively with the data extracted from the treated image.

References

[1] George Nagy. Twenty Years of Document Image Analysis in PAMI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 22(01) :38 – 62, January, 2000.

[2] John D. Hobby. Using shape and layout information to find signatures, text and graphics. *Computer Vision and Image Understanding*, 80(1): 88 – 110, October, 2000.

[3] M. Okada and M. Srigari. Extraction of user entered componentes from a personal bankcheque using morphological subtraction. *International Journal of Pattern Recognition and Artificial Intelligence*, 11(5) :699 – 715, 1997.

[4] M.L. Yu, P.C.K. Kwok, C.H Leung, K.W. Tse. Segmentation and recognition of chinese bank check amounts. *International Journal on Document Analysis and Recognition*, 3(4) :207 – 217, May, 2001.

[5] Nikolay Gorski, Valery Anisimov, Emmanuel Augustin, Olivier Baret, Sergey Maximov. Industrial bank check processing: the a2ia check reader. *International Journal on Document Analysis and Recognition*, 3(4) :196 – 206, May, 2001.

[6] P. Clark and M. Mirhehdi. Combi ning statistical measures to find ima ge text regions. *In ICPR'00*, pages 450 – 453, Barcelona – España, 2000.

[7] S. Djeziri, F. Noboud and R. Plamondon. Extraction of signatures from check background based on a filiformity criterion. *IEEE Transactions on Image Processing*, 07(10) :1425 – 1438, October, 1998.

[8] Mohamed Cheriet. Extraction of handwritten data from noisy grey-level images using a multiscale approach. IJPRAI, 13(5) :665 – 685, June, 1999.

[9] Yong Zhu, Tieniu Tan and Yunhong Wang. Font Recognition Based on Global Texture Analysis. IEEE Trans. On Pattern Analysis and Machine Intelligence. 23(10) :1192 – 1200, October, 2001.

[10] L. Cinque, L. Lombardi and G. Manzini. A multiresolution approach for page segmentation. Pattern Recognition Letters 19(1998) 217 – 225, 1998.

[11] Xyangyun Ye, Mohamed Cheriet, Ching Y. Suen and Ke Liu. Extraction of bank-check items by mathematical morphology. International Journal on Document Analysis and Recognition, 2(2/3) :53 – 66, February 1999.

[12] Xiangyun Ye, Mohamed Cheriet and Ching Y. Suen. A generic system to extract and clean handwritten data from business forms. In Seventh International Workshop on Frontiers in Handwriting Recognition, pages 63 – 72, Amsterdam, 2000.

[13] Hideaki Goto and Hiromoto Aso. Character pattern extraction based on local multilevel thresholding and region growing. In ICPR'00, pages 430 – 433, Barcelona – España, 2000.

[14] Nikolay Gorski, Valery Anisimov, Emmanuel Augustin, Olivier Baret, Sergey Maximov. Industrial bank check processing: the a2ia checkreader. International Journal on Document Analysis and Recognition, 3(4) :196 – 206, May, 2001.

[15] Victor Wu, Rahgavan Manmatha and E. Riseman. Textfinder: An automatic system to detect and recognize text in images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(11) :1224 – 1229, November, 1999.

[16] B. Yu, A. K. Jain and M. Mohiuddin. Address block location on complex mail-pieces. In ICDAR'97, pages 871 – 901, Ulm, Germany, 1997.

[17] F. LeBourgeois and H. Emptoz. Document analysis in gray level and typography extraction using character pattern redundancies. In ICDAR'99, pages 177 – 180, Bangalore – India, 1999.

[18] Minako Sawaki and Norihiro Hagita. Text-line extraction and character recognition of document headlines with graphical designs using complementary similarity measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(10) : 1103 – 1109, October 1998.

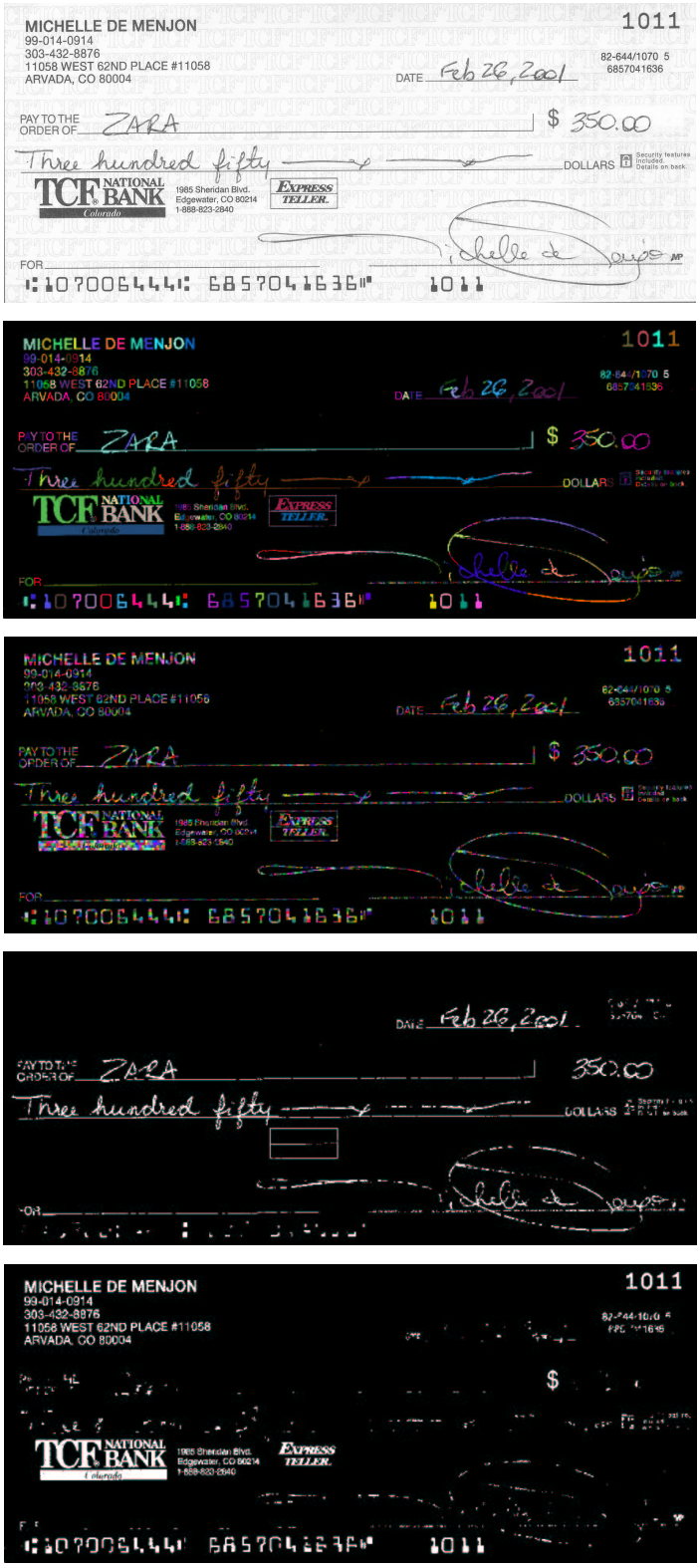
[19] Hiroyuki Hase et al. Character string extraction from color documents. Pattern Recognition, 34(7) : 1349 – 1365, 2001.

[20] J. Liu and Y. Y. Tang. Adaptative image segmentation with distributed behavior-based agents. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(06) :544 – 551, June 1999.

[21] Ke Liu et al. Automatic extraction of baselines and data from check images. International Journal of Pattern Recognition and Artificial Intelligence, 11(04):675 – 697, April 1997

[22] A recursive thresholding technique for image segmentation. IEEE Transactions on Image Processing, 07(06):918 – 921, june 1998.

[23] F. M. Wahl, K. Y. Wong and R. G. Casey. Block segmentation and text extraction in mixed text/image documents. Computer Graphics and Image Processing, 20(2):375 – 390, February 1982.



a)

b)

c)

d)

e)

Figure 2 - a) Original gray level image of an US bank cheque; b) labeled image; c) particioned image; d) handwriting text and e) machine printed text.