Journal of Signal and Information Processing, 2015, 6, 1-8
Published Online February 2015 in SciRes. http://www.scirp.org/journal/jsip
http://dx.doi.org/10.4236/jsip.2015.61001

Extraction of Arabic Handwriting Fields by Forms Matching

Ameur Bensefia

College of Computer and Information Sciences, Al-Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia
Email: abensefia@ccis.imamu.edu.sa

Received 20 December 2014; accepted 10 January 2015; published 21 January 2015

Abstract

Filling forms is one of the most useful and powerful ways to collect information from people in business, education and many other domains. Nowadays, almost everything is computerized. That creates a curtail need for extracting these handwritings from the forms in order to get them into the computer systems and databases. In this paper, we propose an original method that will extract handwritings from two types of forms; bank and administrative form. Our system will take as input any of the two forms already filled. And according to some statistical measures our system will identify the form. The second step is to subtract the filled form from a previously inserted empty form. In order to make the acting easier and faster a Fourier-Melin transform was used to re-orient the forms correctly. This method has been evaluated with 50 handwriting forms (from both types Bank and University) and the results were approximatively 90%.

Keywords

Handwriting Document Analysis, Binarization, Zones of Interest Extraction, Fourier-Mellin Transform

1. Introduction

Nowadays computers have become an integral part of our daily routines, including sending emails, accessing instant information and communicating on social networks. Even with these advanced technologies we still need to deal with paper for so many purposes. Indeed, papers are still an inescapable part of most of business and administrative transactions: such as contracts, application forms, medical forms, checks… etc. The heterogeneity of these types of forms forces us to detect, extract and recognize the handwriting fields from these forms and inject them manually in our respective systems.

How to cite this paper: Bensefia, A. (2015) Extraction of Arabic Handwriting Fields by Forms Matching. *Journal of Signal and Information Processing*, **6**, 1-8. http://dx.doi.org/10.4236/jsip.2015.61001

Thus, many researches have been conducted in this field when the approaches proposed aim to recognize the handwritings text [1]-[4]. These systems take in their input a handwriting sentence or word as an image and provide the electronic representation of the text or the word. However, in real life these handwriting words are often drowned with other representations, draws or graphics.

Our works focus on developing a system which can be considered as a pre-processing system, part of a complete handwriting recognition system. Indeed, the system we proposed aims to take as an input the image of the paper (the form in our case) in real form and then extract the zones of interests, the handwriting fields, to produce their images as an input to the word recognition module "**Figure 1**".

This paper is organized in three main sections: the first one gives a literature review of the zones of interest extractions methods. The second section presents our method which is based on the identification of the form taken as an input of the system first, then extracting the handwriting fields by an original subtraction method. Section 3 and 4 are dedicated to present the databases we use in the evaluation of our system and discuss the performances obtained.

2. Literature Review

In the works of [5] the authors proposed a method to detect in a whole handwriting document all the handwriting numerals (phone number, zip code… etc.). Their system has been built around three main components:

- **Preprocessing:** During this step the text lines are extracted by grouping together the connected components.
- **Connected component labeling:** Since the method was dedicated to the detection of the numerical fields within text lines, they were primarily interested in assigning each connected component to its unknown label which can be either digit or no. Label S has been added for digit separators.
- **Syntactical analysis:** this last stage is crucial for the system as it will verify that some sequences of connected components can be kept as candidates.

Finally the detection of the numerical fields is achieved by analyzing every line of the document. The syntactic analyzer makes its decision in favor of the presence (detection) or absence (reject) of a numerical field on the line under investigation. Syntactic analyzer may propose several locations when a field is detected; the output of the syntactic analyzer is therefore a list of N solutions corresponding to the N best paths found in the observations.

Experiments have been conducted on two non-overlapping databases of handwritten incoming mail documents. The first one (292 documents) has been used to build the learning database and to learn transition probabilities and to parameterize the system. The second one (293 documents) has been used to test the proposed approach "**Table 1**".

Almost the same problem dealt in [5] have been investigated by [6]: Extracting numerical fields in uncon-

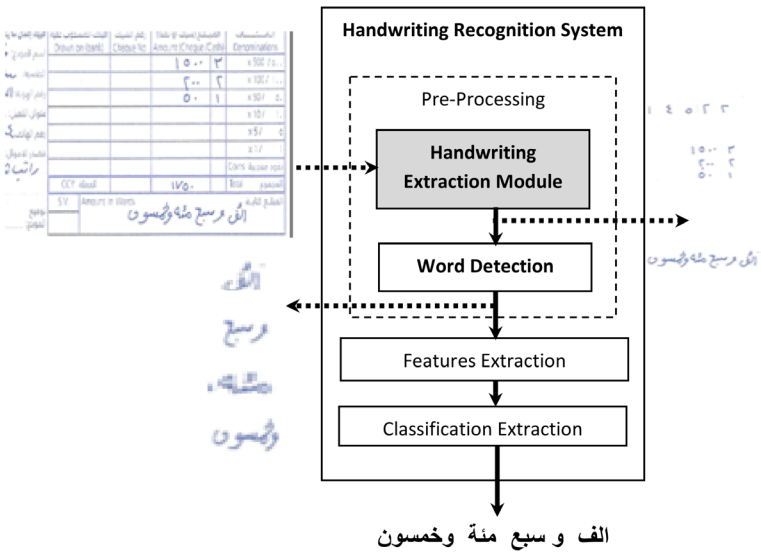

Figure 1. Architecture of a handwriting recognition system.

Table 1. System performance of [5].

	Top 1	Top 2	Top 10
Zip code	41%	66%	88%
Phone number	69%	80%	91%
Customer code	60%	76%	88%

strained handwriting document. They used the incoming mail documents samples which include many kinds of numerical fields (customer reference, phone number, zip code in address…). The approach proposed by the authors begins by:

- **Labeling:** Numerical fields typically contain digits, separators, and touching digits. So four different classes for labeling have been defined: "D" digits, "DD" touching digits, "S" separators, "R" reject.
- **Bounding:** They used a bounding box around each of the connected components.
- **Feature extraction:** "9" contextual feature are calculated based on the bounding box. A Chain code feature set is extracted from the contours of the connected component and a statistical/structural feature set are also calculated (117 features from 6 families).
- **Classification:** a K-Nearest Neighbor classifier (K-NN) has been used to overcome the confusion between digit, separator and double digit classes. A Multilayer Perceptron (MLP) was used to classify the connected components through each of the previous three feature sets.

The works of [7] dealt with the detecting of handwriting in sets of tabular document images that share a common form. An original approach has been adopted. It consists of differentiating pixels in an image using overlapping windows and a voting scheme. An accumulator array, A, is used to store votes at all pixel locations. Handwriting is detected by synchronously passing a window pixel by pixel across both source S and template images T that contains only a blank form and it can be created by taking the point wise median across a set of registered images.

If the pixels in the source vary enough from the pixels in the template, they are marked as handwriting. The comparison between windows is a simple mean difference $d < t$ is set to the median pixel value of the image. If the difference exceeds a threshold ($t = 12$), the weight at each pixel location in A is increased. Conversely, if the difference is below the threshold, the weights are decreased in A for each pixel location in the window.

The accumulator array, A, is equal in size to the source image, and is initialized to 0. Handwriting votes increment values in A, and non-handwriting votes decrement values. When all windows have voted, the array A will contain both positive and negative values. A positive value at an $[x; y]$ location means that the pixel at location $[x; y]$ in the source image is handwriting. Negative values are background. Once the accumulator array is generated, it is used in conjunction with the source image to create a new image, N, with the same dimensions.

The authors in [8] present an automatic extraction of handwritten Arabic components from checks. Two methods were developed: The first one was based on Mathematical Morphology (MM) and the second one was based on Hough Transform (HT).

The numerical amounts are extracted using a Mathematical Morphology (MM) by following these steps:

- The MM filter is applied on the binary image.
- The Otsu method is applied in order to eliminate the background of the initial image.
- Two types of closing filter re applied: one is horizontal and uses the linear structuring element with a length equal to pixels. The second is vertical and uses the linear structuring element.
- Horizontal and vertical filters lead to extracting the straight line existing on the image. The combination of the results of the two stages improves the extraction of the existing box in which the numerical amount is written.
- A mask is constructed and used to identify the numerical amount in the original zone.

To extract the literal amounts, the authors detected and labeled the connected components using horizontal and vertical structuring elements. The authors have also considered the use of the Hough Transform; to extract the literal and numeral amounts. The HT implementation consists in extracting firstly the different feature points then building the different straight segments by choosing a reference point and the gradient orientation at each pixel. After that, the authors applied the HT on the resulted image in order to extract the printed line on which the literal amount and date were written. A labeling process is than initiated in order to identify the literal and

numeral amount.

3. Proposed Method

The system we proposed is built around three different steps: preprocessing, forms identification and finally the handwriting fields extraction:

3.1. Preprocessing

During this stage the input document is preprocessed for the next stages. In our system, the input form is firstly binarized from gray level using the Otsu algorithm [9]. Indeed this algorithm produces an optimal threshold for the binarization by reducing the intra-class variance.

After that, a noise reduction operation is performed using the median filter. Finally a dilation operation may be required to close any gaps appearing in the handwritings.

3.2. Form Identification

In our approach, we decided to work with two different forms "**Figure 2**" a bank form remittance (*form_B*) and a university complaint form (*form_U*). Thus, before extracting the handwriting fields, our system should be able to detect and recognize automatically the form inserted.

This form identification step is based on a unique feature. For the bank form, the unique feature is the bank logo on the top left corner of the form. The university form's unique feature is the university's logo on the top center. The detection of this feature is based on the analysis of histogram horizontal projection of the head of the form. Thus, the presence of a pick black pixels in the center of the histogram means *form_U* form where the presence of this pick in the right means *form_B*.

To recognize these features we need to calculate a histogram of the inserted forms. Some statistical measures are calculated and based on the measures calculated from the empty forms; the system will recognize the input form.

3.3. Fields Extraction

We have observed that instead trying to observe and analyze each single pixel in the forms in order to know if it's belonging to the original ink of the form or belonging to the ink pen used by the writer; we proposed another original approach: despite the fact that all the fulfilled forms are different, they share all a common region (group of pixels) which represent the empty form structure. Indeed, if we want to extract the handwriting fields

(a) (b)

Figure 2. (a) University form; (b) Bank form.

we need to remove the original form structure, in other terms "subtract" from the fulfilled form the empty form.

This method uses a pair of form images, an empty form image and a fulfilled form image. These two images differ only at specific regions. The output of this method is image whose pixel values are the first image minus the corresponding pixel from the second image and then eliminate any part of image that we are not interesting.

However, a rotation step is required to make this operation possible. We used the Fourier-Mellin Transform [10]-[12] to operate this rotation. The use of the same scanner and the same resolution between all the forms can also be required to improve this operation.

The subtraction of the two images is performed in a single pass. The output pixel values are given by the following formula:

$$R(i,j) = \mathrm{Comp}\left[\mathrm{Abs}\left(F_1(i,j) - F_2(i,j)\right)\right] \quad (1)$$

where: $R(i, j)$ a pixel in the resulting image, $F_1(i, j)$ a pixel in the fulfilled form and $F_2(i, j)$ a pixel in the empty form. Comp: means the complement value for the binarized image in the range {0, 255}. Abs: the absolute value of the pixel.

According to this formula only the pixels added to the original form will appear as black pixels. However, if the writer has superimposed his handwriting on the black pixels of the form, these pixels will be lost.

The main benefits of using this technique method are that it is done automatically by the system without any user intervention and it also gives clearer results. The originality of this method is that, in addition to be form independent, there is no extra adjustment to do to extract the handwriting fields, it is also completely independent from the language. Indeed, the writers can write in English, Arabic, French, the method focus only on finding the structure of the original empty form to remove it.

The database we used in this system are made up of two different forms: *form_B* and a *form_U* "**Figure 2**". The idea behind choosing two different forms is to show the flexibility and the independency of our system form any kind of form.

The forms chosen can be fulfilled either in Arabic, English or any other language, because the method we proposed is completely language independent. From each of the selected forms we collect 25 copies and asked some students to fill them in their own natural writing style. The administrative forms are an A4 paper size where the bank form size is A5. The bank form was taken from a national bank in Saudi Arabia; it is a deposit form. The administrative form is a student department complaints form.

All these forms contains numeric as well as alphabetical characters, see some samples in "**Figure 3**". They have been scanned in a grayscale image color 300 dpi.

4. Experimentations and Results

The evaluation of the proposed method has been achieved in two steps:

4.1. Evaluation of the forms Identification

We presented to our system all the 50 fulfilled forms from both groups (*form_U* or *form_B*) in a random order and we counted the number of times when the system proposed the form name correctly. All the forms used have been detected correctly in 100% of the cases.

4.2. Evaluation of the Handwriting Fields Extraction

In order to evaluate this method and establish a significant performance rate that can be analyzed and discussed; we decided to proceeded by counting first the number of handwriting words, digits or numbers present in each group of forms than we counted their number in the output of the system if they have been correctly and completely extracted "**Table 2**".

The performance of our method reaches 84% with the *form_U* when this rate increases to 92% with the *form_B*. some results is presented in "**Figure 4**".

The results obtained by our whole system (detection and extraction) are very promising. Regarding the forms detection theses results were expected since the parameters used to distinguish the two forms don't contain any intersection or confusion range between them. However, it seems hard to ensure that this rate will stay so high if we used more than 2 forms particularly if these forms present a high degree of similarity.

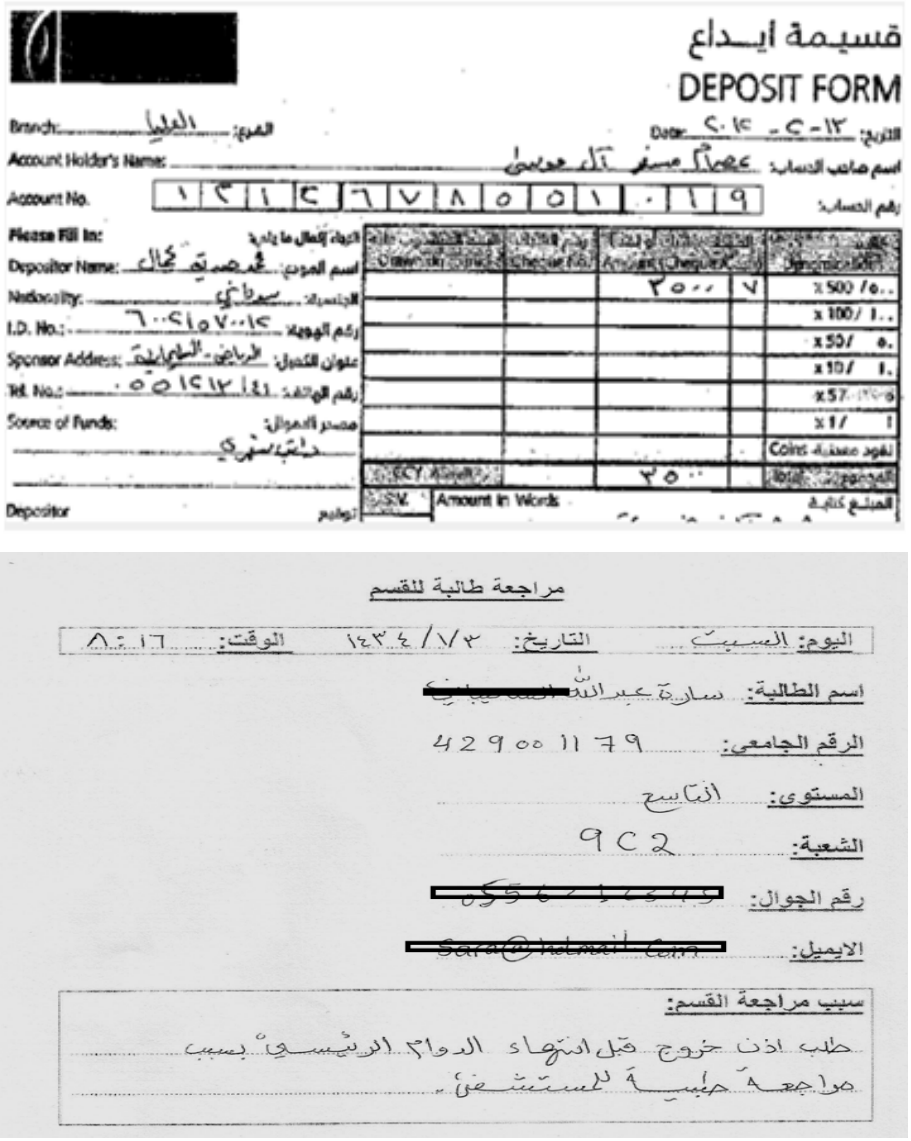

Figure 3. Samples of the fulfilled forms.

Table 2. Our system performance based on the 2 forms.

Forms	Number of words correctly extracted/ total number of words	Rate
Bank form	685/743	92.19%
University form	211/251	84.06%

Regarding the extraction method, the results were very promising and positive since the global performance of the method for the 2 forms is around 90%. These results don't take into account the quality of the text extracted. Indeed, some additional post-processing methods can be applied such as dilatation or noise removal in order to make these fields more exploitable by the word recognition module.

5. Conclusions

In this paper we present a handwriting extraction fields system, which can be seen as a preprocessing module

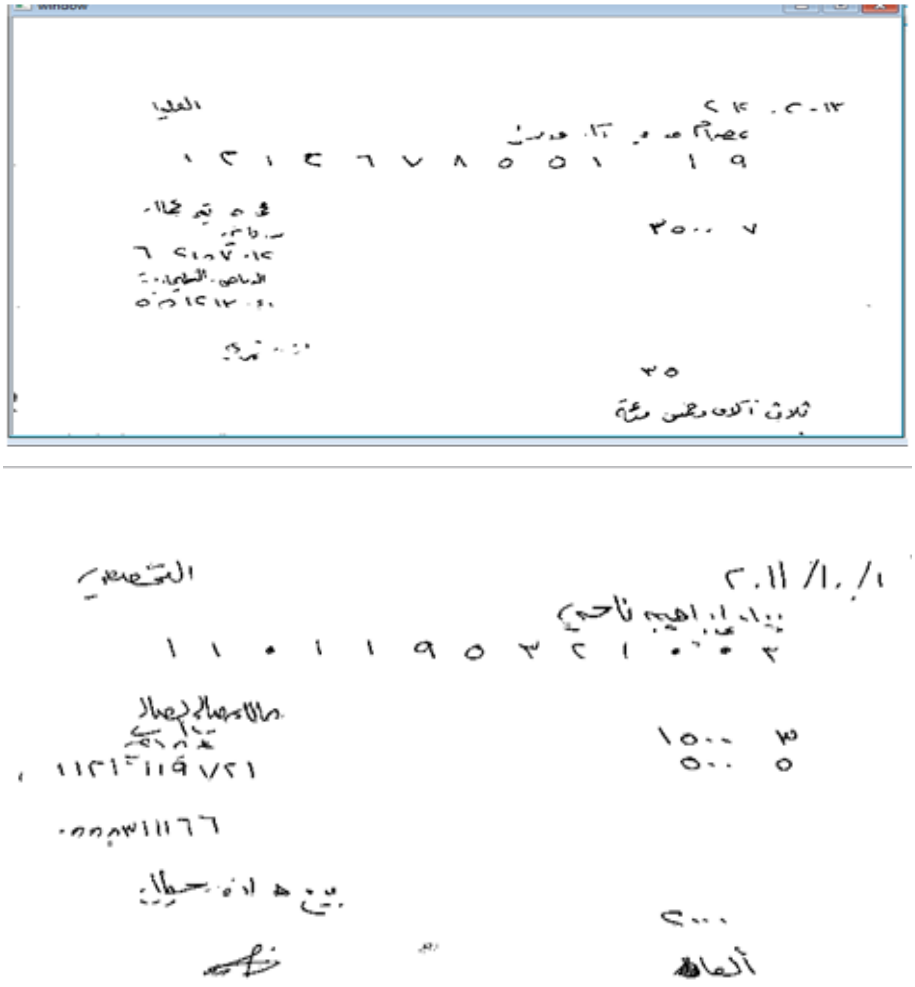

Figure 4. Some samples of the extracted handwriting fields.

in a complete handwriting recognition system. This module extracts the zone of interests (handwriting fields) in any entered form and makes these zones available to the recognition module.

We propose a free language and a free form method. Indeed, this method remove the original form structure, in other terms “subtract” from the fulfilled form the empty one.

To make this operation possible a preprocessing step was applied, we binarized all the forms using the Otsu algorithm then we removed some noise. The second step was the form identification since we evaluated our system with 2 different forms types and then an orientation step was necessary to make the matching operation possible; to this purpose we used a Fourier-Mellin transform.

We evaluated our system with 50 forms belonging to two different forms (bank form and a university form). The global performance of our method gives around 90% of good extraction.

We propose in the future, to investigate extraction of the handwriting fields from any form or document without any a priori about the structure of this document, which is once again more close to the reality.

References

[1] Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H. and Schmidhuber, J. (2009) A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 855-868.

[2] Lorigo, L.M. and Govindaraju, V. (2006) Offline Arabic Handwriting Recognition: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**, 712-724. http://dx.doi.org/10.1109/TPAMI.2006.102

[3] Plamondon, R. and Srihari, S.N. (2000) Online and Off-Line Handwriting Recognition: A Comprehensive Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 63-84. http://dx.doi.org/10.1109/34.824821

[4] Senior, A.W. and Robinson, A.J. (1998) An Off-Line Cursive Handwriting Recognition System. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 309-321. http://dx.doi.org/10.1109/34.667887

[5] Koch, G., Heutte, L. and Paquet, T. (2003) Numerical Sequence Extraction in Handwritten Incoming Mail Documents. *Seventh International Conference on Document Analysis and Recognition*, **1**, 369-373.

[6] Chatelain, C., Heutte, L. and Paquet, T. (2004) A Syntax-Directed Method for Numerical Field Extraction Using Classifier Combination. 9*th International Workshop on Frontiers in Handwriting Recognition IWFHR*-9, 26-29 October 2004, 93-98.

[7] Clawson, R. and Barrett, W. (2012) Extraction of Handwriting in Tabular Document Images. Family History Technology Workshop at Rootstech.

[8] Samoud, F.B., Maddouri, S.S., Abed, H.E. and Ellouze, N. (2008) Comparison of Two Handwritten Arabic Zones Extraction Methods of Complex Documents. *Proceedings of International Arab Conference on Information Technology*, Hammamet, 1-7.

[9] Otsu. N. (1979) A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man and Cybernetics*, **9**, 62-66. http://dx.doi.org/10.1109/TSMC.1979.4310076

[10] Adam, S., Rousseau, F., Ogier, J.M., Cariou, C., Mullot, R., Labiche, J. and Gardes, J. (2001) A Multi-Scale and Multi-Orientation Recognition Technique Applied to Document Interpretation Application to French Telephone Network Maps. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, **3**, 1509-1512.

[11] Liu, Q., Zhu, H.Q. and Li, Q. (2011) Object Recognition by Combined Invariants of Orthogonal Fourier-Mellin moments. 8*th International Conference on Information, Communications and Signal Processing* (*ICICS*), Singapore, 13-16 December 2011, 1-5.

[12] Sharma, V.D. (2010) Generalized Two-Dimensional Fourier-Mellin Transform and Pattern Recognition. 3*rd International Conference on Emerging Trends in Engineering and Technology* (*ICETET*), Goa, 19-21 November 2010, 476-481.