# FORM IDENTIFICTION AND SKEW DETECTION FROM PROJECTIONS

N. Liolios N. Fakotakis and G. Kokkinakis

Wire Communications Laboratory
University of Patras
Patra 26500, GREECE

## ABSTRACT

In this paper we describe a system we have built to solve the preprinted forms identification and field extraction problem for Optical Character Recognition (OCR) applications. The strength of this system is that unlike other approaches it solves the problem in the most general and unrestricted sense. It works equally well for any type of preprinted form because it does not rely on any special features like patterns of line crossings or other symbols found only in a particular type of form. We have used the power spectrum as a shift invariant feature vector of the form's horizontal projection from which we identify the type of form and detect rotation. The horizontal and vertical projections themselves are also used to detect the shift of the form. Unlike the expected loss in response time to the benefit of generality, the proposed system is fast, highly accurate, even at reduced resolutions and with minimal user intervention it can be trained to recognize new types of forms.

## 1. INTRODUCTION

Office automation is steadily decreasing the number of documents that contain handwritten parts but companies are still processing a vast amount of documents with conventional methods. Even companies that are in the process of becoming fully Office-Automated have a need for a system that will somehow convert their old documents to some kind of electronic form [1].

Methods and tools have been developed in the past [2] that to a satisfactory degree solve the problem of printed character recognition and to a less than satisfactory degree of handwritten character recognition. The problem of optical character recognition inherits added complexity when the document to be processed is a preprinted form, with fields that are initially blank and are then filled by a customer. The most representative example in this case is an insurance application form. The user-supplied information can be in either handwritten or machine typed form. To make things worse the document might be contaminated with noise and may have lost resolution during a fax transmission, it may have been shifted, skewed or deformed in several possible ways.

The form to be processed contains two types of information: The preprinted portion of the form and the customer supplied data. Before the preprinted portion can be separated from the user filled information the type of form has to be identified. The system may be possibly trained to recognize many different types of forms that are used in an entire organization. During the identification process the system has to determine which one of the stored prototypes (blank forms) matches best the incoming one. Therefore before the fields (user filled information) can be extracted the form may have to be rotated and possibly shifted to exactly match the prototype it is identified with. Only after this match for rotation and shift has been achieved can the original field locations be used as a template in order to extract the corresponding information from the processed form.

There are several generic algorithms that can be used to detect skew [3],[4] and shift (translation) but all are prohibitively expensive as far as computation time is concerned. Thus a system based on these algorithms could handle any type of form but it would be unusable as a commercial product.

Form Identification existing today, rely on one of the three methods that are described briefly below:

- **Use of special symbols or structures on the form**. These symbols are initially located on the form and then compared to a prototype to determine the degree of rotation or translation. This approach requires a specially designed form and therefore it is not suitable for processing existing types of forms.
- **Detection of vertical and horizontal lines** from which the skew and shift can be determined [5]. This method is computationally very intensive because it uses the Hough transform for line detection. Some improvements to this method have been made but they require user intervention to determine the subparts of the form that contain the lines of interest when the initial form is presented to the system for training.
- **Special patterns of line crossings** can be used, the location of which, when detected, can determine the amount of skew and translation to be corrected [6]. The patterns or line crossings can be used as a vector for form identification as well . This system does not require any user intervention during training but it does rely on the form having these line crossings.

Our approach does not have the limitations described above, which makes it more generic and useful with any type of preprinted form. The advantage comes from the fact that we start

[·] This work is supported by the European Commission Telematics Applications Program, project LE-1 1802 ACCeSS.

with the form's horizontal projection, which is unique enough for a large set of form types and styles and then we extract the power spectrum of it as a shift invariant feature vector. The power spectrum is subsequently used for the identification of form type and the rotation detection. The horizontal and vertical projections are then used unmodified to detect the horizontal and vertical shift. Furthermore the implementation of these algorithms has resulted in a fast system with adequate response time.

This paper is organized as follows. In Section 2 we describe the system. In Section 3 we analyze our experimental results and finally in Section 4 we draw some conclusions and discuss our plans for future work.

## 2. SYSTEM DESCRIPTION

The preprinted forms identification module of our system (designed in the framework of the LE-1 1802 project : ACCeSS) is the preprocessor of the handwritten text recognition part of the system. The modem receives one of the known forms that the system is trained with, as either a scanned or faxed file. After the type of form is identified, the form is deskewed and translated properly so that the parts filled by the user are located and their contents extracted.

Although not initially specified, we decided to extend the design in such a way so that it can be a generic form processing tool for all types of forms. We also decided that such a tool in order to be usable, and it should require the least possible human intervention and should run in real time.

For the field definition part we build a tool with a graphical interface that the user employs to specify the blank fields of interest where the handwritten text is expected. The contents of these fields have to be extracted and forwarded to the handwritten text recognition module of the system. In order to achieve this a clean deskewed form is loaded as the background and rectangles are drawn over the defined areas of interest which is where the information of interest is to be located when the filled form of this type has been corrected for rotation and shift.

### 2.1 Feature Extraction

To solve the problem of identifying the form without making use of any special symbols or line crossing patterns, we decided to use the blank form's vertical and horizontal projections as the feature vectors from which the form would have to be identified.

Initial tests showed that the projections can be used successfully for form identification providing there is no rotation or shift. Part of the system speed is gained by the use of a specially designed projection algorithm which processes the image memory in a serial serially, avoiding coordinate arithmetic completely while obtaining both horizontal and vertical projections simultaneously.

In order to solve the problem of correctly identifying a rotated and possibly shifted form we decided to rotate the blank prototype to a number of pre-specified angles and to obtain a horizontal projection for every different rotation angle. At this stage about 60 different projections are generated for each form for a -30° to +30° rotation at 0.1° steps. The projections are stored in a database using the form name, the rotation angle and the projection type (Vertical or Horizontal) as the key for subsequent retrievals.

The vertical projections are obtained only for 0° degrees of rotation since they do not contain any significant information about a document's skew angle in portrait orientation. The vertical projections however are used for horizontal shift detection in the same manner as horizontal projections are used for vertical shift detection.

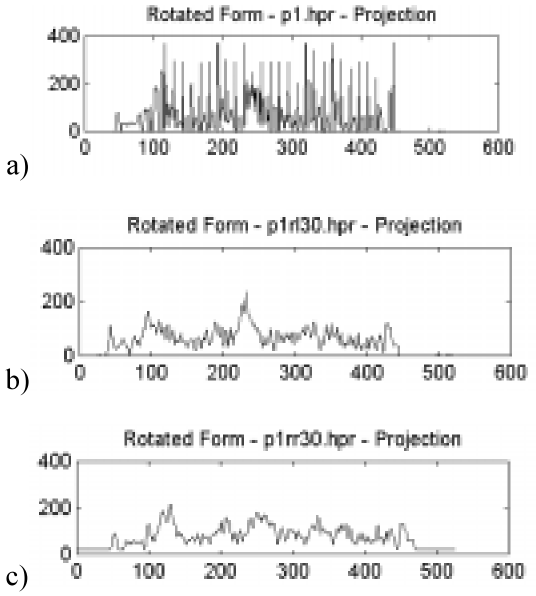

**Figure 1**. Projections of the same form at different angles: (a) at 0°, (b) at -30°, (c) at +30°

Figures 1.a, 1.b and 1.c show the horizontal projections of a blank form for 0°, -30° and +30° of rotation. It turns out that there is enough variance in the projections to successfully identify the degree of rotation once the type of the form has been identified.

To identify a possibly shifted form from its projection we should store every possible shifted projection for every rotation angle as a prototype for training. Alternatively the projection of the incoming form could be shifted in all possible ways and each one of them would have to be compared with the stored unshifted projections to find the best match. Both of these methods were tested and the recognition results were very satisfactory. Both methods however require a significant amount of computation and disk I/O time, which renders them unusable.

It is therefore evident that a shift invariant transformation is required for the projection. We decided to use the power spectrum of the horizontal projection based on the short time DFT with 50% overlapping Hanning windows and a 0.95 confidence interval. Figures 2.a, 2.b and 2.c show how pronounced the horizontal projection differences are on a form for a 0, -200 and +300 pixel shift.

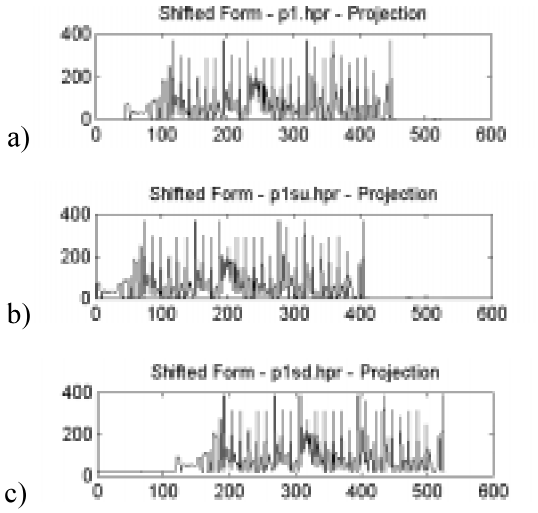

**Figure 2**. Projections of the same form with different shifts: (a) No shift, (b) -200 pixel shift, (c) +300 pixel shift.

Figures 3.a, 3.b and 3.c show the power spectra obtained for the projections in figures 2.a, 2.b and 2.c correspondingly. The similarity of the three parts of figure 3 indicate that the power spectrum forms a feature vector that can be used for form identification in a shift invariant manner.

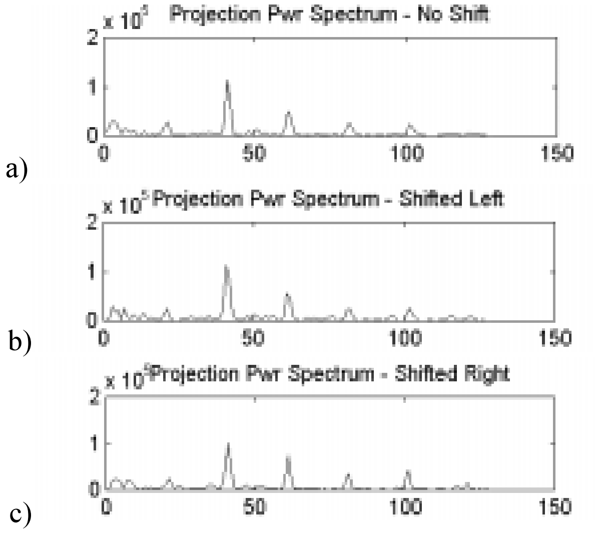

**Figure 3**. The power spectra of the same projection but with different shift amounts: (a) No shift, (b) -200 pixel shift, (c) +300 pixel shift.

## 2.2 Training and Type Identification of the Form Type

We store the horizontal projection along with the power spectrum for all the rotation angles, as specified earlier, for every form that the system has to learn.

Vector quantization (VQ) and hidden Markov models (HMMs) were used for system training and form type identification. All power spectrum vectors that resulted for every rotation angle of the same form were assigned to the same class labeled by the form's ID. The centroid of the class was subsequently used for training one HMM for every class (form). This set of HMMs was sufficient to solve the form type identification problem successfully.

## 2.3 Skew and Shift detection

Once the type of form is identified, the rotation angle as well as the amount of horizontal and vertical shift have yet to be determined. Since the horizontal projections were generated for all predefined rotation angles we used (again) a second set of HMMs, one HMM per form type, which was trained by the vector resulting from the horizontal projection of the form at that angle. In this case no vector quantization was necessary since every projection is a class in itself, labeled by the corresponding rotation angle.

With the form deskewed the problem is reduced to determining the horizontal and/or vertical shift. The shift detection problem is solved using direct Euclidean distance as a measure of similarity between any two projections. By performing continuous shifts on the projection of the form being processed until the shift that results in minimum distance is found. The horizontal projection is used for vertical shift detection and the vertical one to detect the horizontal shift.

To reduce computation we do not shift the projection but a positive or negative index to the projection vector. To reduce computation further the search for the best shift is done in three stages:

1. Find the best match using a shift step of 50 pixels.
2. Search for a better match in the neighborhood of -50 to +50 pixels from the previous match using a shift step of 10
3. Change the neighborhood to an area -10 to +10 and repeat the search around the previous best match with a shift step of 1.

Again, after the horizontal and vertical shifts are determined in the same manner instead of shifting the whole form the prototype form's field coordinates are shifted in the opposite direction.

# 3. EXPERIMENTAL RESULTS

The system was trained using 26 blank form prototypes and tested with a set of 300 forms that contained a variable amount of handwritten text in the blank fields. Significant amount of noise was also introduced artificially during scanning. The system was tested for recognition accuracy at different scanner resolutions and the results are summarized in Table 1.

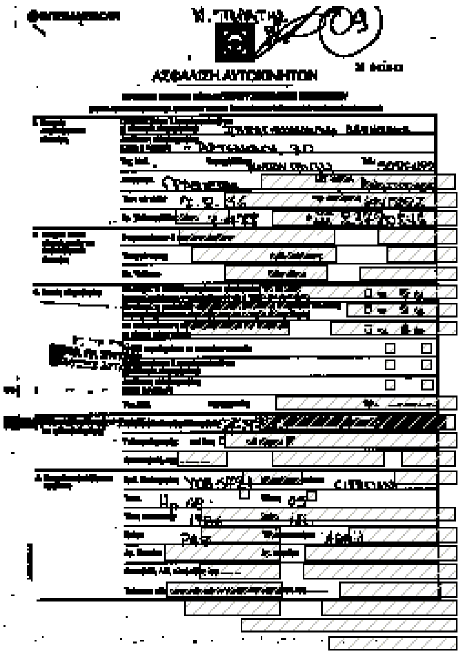

**Figure 4.** Initial positioning of fields after identification.

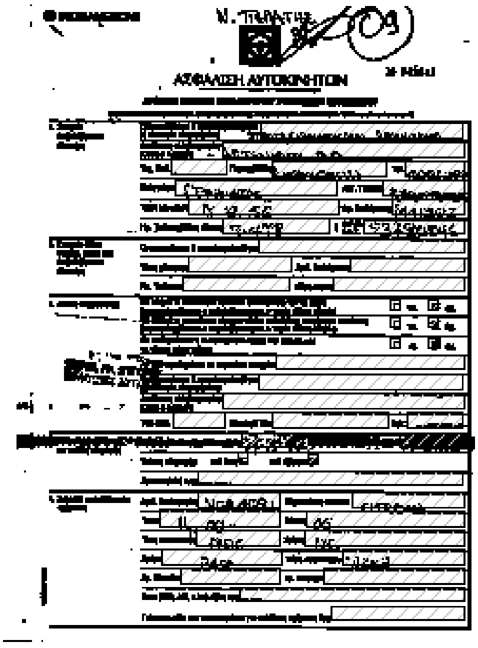

**Figure 5.** Position of fields after skew and shift correction.

Figure 4 is a section of a form put in the system, which shows how misplaced the fields are after identification. This particular form was shifted up and left during scanning. A significant amount of noise can be seen as a black smudge near the center of the form. The rotation is not quite visible but nevertheless the system detected a skew of $0.4^0$ the left shift 74 pixels and the upshift 237 pixels. Figure 5 shows how well the fields are positioned after deskewing and deshifting.

| Dpi | Correctly Identified (%) | CPU Time (Sec) |
|---|---|---|
| 300 | **100** | **4.9** |
| 150 | **100** | **2.3** |
| 75 | **93.1** | **1.8** |
| 35 | **85.1** | **1.2** |

**Table 1.** Resolution vs. Recognition Rate and CPU Time.

The CPU time was measured on a 200 MHz Pendium CPU running Windows NT Server V4.0.

## 4. CONCLUSIONS AND FUTURE WORK

We proposed a system that faces the form type identification and field extraction problem in the most generic case. It turns out that the power spectrum is an excellent choice as a shift invariant feature vector for accurate identification of the form type. The choice of shift invariant feature vectors contributes highly to the system's fast response time.

In the near future we plan to incorporate a document defect models to create artificial data to test extreme cases as well as large numbers of documents.

The system described in this paper can easily be extended to work on a form in either landscape or portrait orientation. An obvious solution would be to obtain vertical projection for every angle and test them for best fit as well.

## 5. REFERENCES

[1] Andreas Dengel, Rainer Bleisinger, Rainer Hock, Frank Fein,and Frank Hones, "*From Paper to Office Document Standard Representation*", Computer, Vol. 25, No. 7, July 1992, pp. 63-67.

[2] Shunji Mori, Ching Y. Suen and Kazuhico Yamamoto, "*Historical Review of OCR Research and Development*", Proc. IEEE, Vol. 80, No. 7, July 1992, pp. 1029-1058.

[3] Henry S. Baird, "*The Skew Angle of Printed Documents*", Proc. Conf. Of the Society of Photographic Scientists and Engineers, 1987, pp. 14-21.

[4] Stuart C. Hinds, James L. Fisher, and Donald P. D'Amato, *A Document Skew Detection Method Using Run-Length Encoding and the Hough Transform*", Proc. 10th Int'l Conf. Pattern Recognition, 1990, pp. 464-468.

[5] Richard Casey, David Ferguson, K. Mohiuddin, and Eugene Walace. "*Intelligent Forms Processing System*", Machine Vision and Applications, Vol. 5, 1992, pp. 143-155.

[6] Suzanne Liebowitz Taylor, Richard Fridzson, and Jon A. Pastor. "*Extraction of Data from Preprinted Forms*", Machine Vision and Applications, Vol. 5, 1992, pp. 211-222.