Information Extraction from Historical Semi-Structured Handwritten Documents

Xujun Peng
Raytheon BBN Technologies
10 Moulton St, Cambridge, MA 01801
xpeng@bbn.com

Huaigu Cao
Raytheon BBN Technologies
10 Moulton St, Cambridge, MA 01801
hcao@bbn.com

Krishna Subramanian
Raytheon BBN Technologies
10 Moulton St, Cambridge, MA 01801
ksubrama@bbn.com

Elizabeth Boschee
Raytheon BBN Technologies
10 Moulton St, Cambridge, MA 01801
eboschee@bbn.com

Rohit Prasad
Raytheon BBN Technologies
10 Moulton St, Cambridge, MA 01801
rprasad@bbn.com

Prem Natarajan
Raytheon BBN Technologies
10 Moulton St, Cambridge, MA 01801
pnataraj@bbn.com

ABSTRACT

In this paper, we describe our approach to extract salient events such as birth and death records from historical French parish documents that contain free-form handwritten text. The challenges posed by these documents to the current state of the art in handwriting recognition and information extraction go well beyond the generic challenges in recognizing handwritten text such as style variations, irregular baselines, poor legibility, etc. Our approach for extracting salient events from such documents has the following processing steps: (1) pre-processing for noise removal and high-quality binarization, (2) OCR for text recognition, and (3) statistical information extraction for event record extraction. In this paper, we focus on preprocessing techniques for robust binarization in presence of different types of degradations that are common in historical documents. We provide a detailed description of our system, experimental setup, and results for each stage of the processing. In addition, we compare different approaches for preprocessing by assessing their impact on OCR performance.

General Terms

Application

Keywords

Historical document, preprocessing, OCR

[1]The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

[2]Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

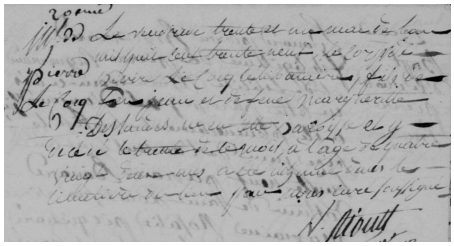


Figure 1: An example of degraded historical document.

1. INTRODUCTION

Historical documents contain important information spanning cultural, genealogy, and scientific information. However, these documents are extremely challenging for analyzing the layout and content. These documents suffer from degradations such as uneven illumination, smear, ink bleedthrough, etc. In Fig.1, we show a sample document from French Parish collection. On such documents, existing preprocessing techniques fail to deal with such artifacts.

Most document analysis and recognition systems that use Optical Character Recognition (OCR) rely on high quality binarized images. However, techniques such as OTSU [6], which use a global threshold for binarization typically break on historical documents. While several adaptive binarization methods have been proposed in the past [5, 8], they often result in a different set of processing aritifacts. Ink-bleed is another serious problem for historical document images that makes binarization an extremely difficult task. Misclassification of bleed ink as foreground text can lead to the failure of text extraction, text line detection and OCR consequently. In this paper, we focus on a novel binarization technique that uses the stroke directional characteristic of cursive style handwritings. Specifically, we apply a nonlinear normalization procedure before binarization to overcome the uneven illumination problem and use the orientation information strokes to remove the bleed ink from degraded historical document image.

The rest of the paper is organized follows. Section 2 de-

scribes the binarization where the original document image is normalized by using a non-linear mapping. Section 3 discusses the method for bleed-through ink removal. Other preprocessing steps such as text line slant correction and height estimation are covered in section 4. The OCR system and information extraction system are described in section 5 which is followed by section 6 of the experimental setup and results; the conclusions are described in section 7

2. BINARIZATION

Based on the assumption that the background intensity is consistent within the local area of document image, we calculate the mean value $\mu_{x,y}$ and the variance $\delta_{x.y}$ over a small $m \times n$ sized window for every pixel (x, y). Normally, if a pixel (x, y) is from the text area, its variance $\delta_{x,y}$ is large. Otherwise, its variance is small if it is a pixel from pure background areas. Then, we use a logistic function to map the intensity of each pixel to a new domain according to its variance which is described by Eq. 1:

$$\mathcal{S}(x,y) = \mu_{x,y} \left\{ \frac{1-k}{\left(1 + e^{-B\left(\frac{\delta_{x,y}-\delta_{min}}{\delta_{max}-\delta_{min}} - M\right)}\right)^{1/\nu}} + k \right\} \quad (1)$$

where B is the parameter which controls the growth rate of the logistic curve, M and nu control the time maximum growth occurs and k is the minimum value the curve can achieve. δ_{min} and δ_{max} in the equation are minimum variance and maximum variance over the entire document image [7].

By applying the above non-linear mapping, the original tensity of a given pixel is mapped close to the local mean value if this pixel is from text areas, otherwise the mapped intensity is much lower than the local mean value. Thus, by comparing the mapped intensity and the original intensity, the initial binarization can be carried out according to the following equation:

$$I'(x,y) = \begin{cases} 255 & \text{if } I(x,y) > S(x,y) \\ 0 & \text{if } I(x,y) < S(x,y) \end{cases} \quad (2)$$

3. BLEED REMOVAL

Cursive handwritings often exhibits a slant which causes the ink bleeds from the other side of the page to have inverted slant from the actual foreground text. Based on this important observation, we formulate a Gabor filter based stroke orientation detection method which detects the ink-bleed from the reverse side of the historical document image.

Gabor filters are band-pass filters which are modulations of a complex sinusoidal and Gaussian function and can be used to detect the direction of strokes. The details of Gabor filters can be found in reference [2]. In our work, we compute the Gabor responses for every pixel with different directions and the maximum response associated with its corresponding angle is obtained for each pixel which are expressed by Eq. 3 and Eq. 4:

$$R(x,y) = \max_{\theta} G_{\theta}(x,y) \quad (3)$$

$$\hat{\theta}(x,y) = \arg\max_{\theta} G_{\theta}(x,y) \quad (4)$$

In our experiments, we traverse the angle of each Gabor filter from 0° to 180°. To remove long bleed inks from the reverse side of the document image whose stroke angles are in the range of 90° to 180°, we measure the differences between actual stroke and bleed ink using the Gabor response and sine value of the angle of each pixel:

$$d(x,y) = sin(2\theta) \times R'(x,y) \quad (5)$$

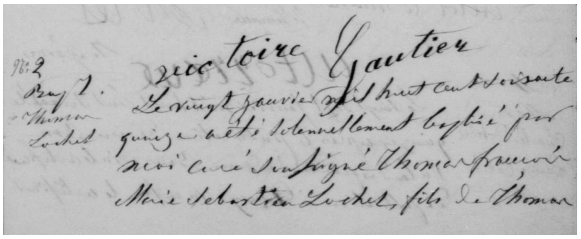
where $R'(x, y)$ is the Gabor responses for each pixel after a linear mapping over entire document image. From Eq. 5 we can see that the pixels on long bleed ink from the reverse side of the document image have a large negative $d(x, y)$ values, and the actual long strokes have large positive $d(x, y)$ values.

If we directly use $d(x, y)$ to filter out bleed inks, it is easy to break the actual strokes into several pieces, especially for circles and curves in the historical document image. Thus, we feed $d(x, y)$ and the intensity value $I'(x, y)$ of each pixel after initial binarization into a conditional random field (CRF) where $d(x, y)$ and $I'(x, y)$ are considered as values for associated potentials for each pixel, and the differences of $d(x, y)$ between neighboring pixels are considered as interaction potentials for each pixel. The optimal solution of the conditional random field which corresponds to the final binarized image is obtained by using a graph cut algorithm which is described in [1]. More details of conditional random field can also be found in the same reference and [3].

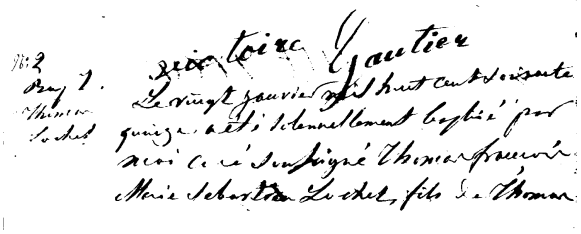
In Fig. 2, we show the results of binarization by using the proposed method on degraded historical document image where Fig. 2(a) is the original degraded image, Fig. 2(b) shows the binarization result using global thresholding method, Fig. 2(c) is the initial binariation result after nonlinear mapping and Fig. 2(d) shows the binarization result using bleed-through ink removal and CRF method proposed in this paper. We can see that by considering the orientation information of the strokes, one can effectively remove the bleed ink from reverse side of the document image and keep the actual strokes by using nonlinear mapping.

4. SLANT CORRECTION

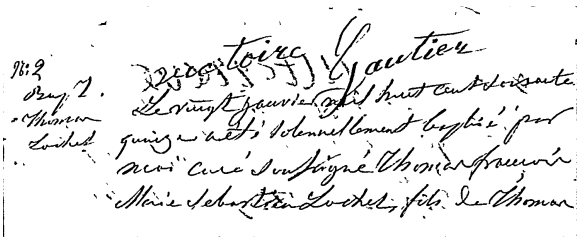
Prior to optical character recognition (OCR), we extract text lines from the binarized document images and up-sample the line images from 300 dpi to 600 dpi. In addition, we correct for slant in foreground text that is common in cursive handwriting. In particular, the writings in the French Parish corpus exhibit significant slant. In order to correct for slant, we applied an affine transformation that can be expressed as:



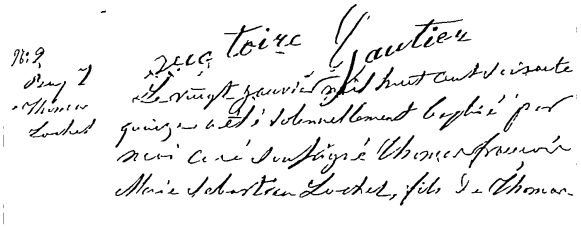
(a) Original image



(b) Binarization using single threshold



(c) Initial binarization with nonlinear mapping



(d) Final binarization using bleed ink removal and CRF

Figure 2: Examples of binarization.

(a) Line image before slant correction

(b) Line image after slant correction

Figure 3: Examples of slant correction.

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (6)$$

where x and y are locations of the original pixel and x' and y' represent the new location of the pixel after affine transformation. To the slant correction, we can reduce the affine transformation to a shear transformation which has the transformation matrix described as: $\begin{bmatrix} 1 & \lambda & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, where λ controls the degree of shear transformation. In our work, we choose $\lambda = \tan(-30^\circ)$. In Fig. 3, we show an example of slant correction for a text line image where we can see that after slant correction, the vertical overlapped areas of neighboring characters are reduced enormously which benefits the latter OCR system.

To further decrease the effect of vertical overlapping between neighboring characters, we estimate the height of the text line and crop the text line image according this height. After cropping, only the main part of each character is retained and those long strokes vertically overlapped with other characters are removed. The procedure of this cropping can be described as:

1. Calculate the mean value μ and variance σ of the height of connected components in the line image;
2. Compute the horizontal profile of the line image using equation: $h(y) = \sum_x (255 - I'(x, y))/255$, where $h(y)$ counts the number foreground (text) pixels for a given line y;
3. Use a sliding window whose size is $\mu + 2\sigma$ on the profile of line image to find out the best location of the text line, where the maximum number of foreground pixels are included.
4. Crop the text line image according to the best location of the sliding window.

5. OCR & IE SYSTEM

5.1 Configuration of OCR System

In this subsection, we discuss the configuration of the OCR system we used which is based on the BBN Byblos OCR system presented in [4]. The Byblos OCR system models handwritten text as the output of hidden Markov model (HMM) based character models.

Prior to training and recognition phase, features are extracted for each text line image by dividing the entire line image into a sequence of thin overlapped vertical frames. The extracted features include percentiles of intensity values, angle, correlation, and energy of each frame. Linear Discriminant analysis (LDA) is then implemented to reduce the feature dimension from 33 to 15.

In the training phase, a left-to-right HMM with 14 hidden states is used to model each individual character. Each state of the HMM has an output probability distribution over the features modeled as a Gaussian mixture model. During the recognition phase, the recognition engine performs a two-pass decoding using glyph model and the language model. In our experiments, a trigram language model is trained using 100k French words. The first round of decoding is forward decoding which uses a fast search scheme and a bi-gram language model to provide a restrict subset of hypothesis. The second round of decoding uses backward search scheme and a trigram language model on the output of the first forward pass to obtain a N-best list of hypothesis. The final recognition results are obtained by using a re-ranked combination of the acoustic score and a language model score which is tuned on a development set.

5.2 Information Extraction System

Based on the output from OCR system, we extract information such as event (marriage, burial and baptism) type, event date, person's name, etc. Our information extraction (IE) system has two main components. In the first component, we use BBN's natural language processing tool IDX to perform sentence chunking, token identification, name-finding, and date-finding on the transcribed text of the images. The second component then uses this information to extract the target events, combining the output of IDX with contextual cues and weighted sets of patterns to identify the persons involved and their roles, as well as event dates. The system we evaluated here is the initial prototype for this task; further improvements are expected on all fronts.

6. EXPERIMENTAL RESULTS

6.1 Corpus description

In our experiments, we collected 800 historical French parish documents that contain free-form handwritten text. The data consists of scanned images of handwritten records of birth, death, marriage, as well as their corresponding ground truth annotations. The document images suffer several degradations such as uneven illumination, stains, bleed-through ink from reverse side of image, smear strokes, etc. The images were scanned through a fair quality scanner, at a resolution of 300 dpi. The ground truth annotations included the co-ordinates of bounding boxes around each text line and the corresponding tokenized transcriptions.

6.2 Experiment setup and results

From the total of 800 document images, we randomly selected 700 pages for training. The remaining 100 images were split equally into a tuning and a validation set. Prior to training and recognition, the preprocessing procedures described in previous sections were implemented which included binarization on entire document image, slant correction and cropping on line image.

Experiment #	Description	Word Error Rate (WER)
1	Proposed binarization	75.7%
2	Up-sampling	74.5%
3	Slant correction	70.3%
4	Final system	66.5%

Table 1: Results of OCR.

The OCR engine was trained using BBN Byblos system where the glyph model was trained on the collected line images and the language model was trained on the French Gigaword corpus and collected French parish transcriptions. The decoding lexicon consisted of 100k of the most frequent words in French Gigaword corpus plus 5k words in French parish corpus. The out of vocabulary (OOV) rate of the test set measured against the 105K lexicon is 4.64%.

In our experiments, we measured the impact of different preprocessing methods on OCR performance. Table 1 shows the performance in terms of word error rate (WER) on the test set where the base system (Experiment 1) is carried out using the proposed binarization and bleed ink removal method and final system is the system with all proposed preprocessing steps which include binarization, up-sampling, slant correction and line image cropping. From table 1 we can see that, by applying slant correction and line image cropping, the WER decreases around 8%.

7. CONCLUSIONS

In this paper, we proposed a new binarization method for historical free-style document images which uses non-linear mapping scheme and stroke orientation information to overcome the problem of uneven illumination and bleed-through. In order to increase the OCR performance, other pre-processing steps, including up-sampling, slant correction and line image cropping, are also described in our paper. From the experiments, we can see that the proposed preprocessing methods can effectively improve the performance of the OCR system on historical semi-structured handwritten document images and facilitate the higher level information retrieval.

8. REFERENCES

[1] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(9):1124 –1137, sept. 2004.

[2] S. E. Grigorescu, N. Petkov, and P. Kruizinga. Comparison of texture features based on gabor filters. *IEEE Trans on Image Processing*, 11(10):1160–1167, 2002.

[3] J. Lafferty, A. Macullum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequential data. In *8th International Conference on Machine Learning (ICML)*, pages 282–289, 2001.

[4] P. Natarajan, S. Saleem, R. Prasad, E. MacRostie, and K. Subramanian. Multi-lingual offline handwriting recognition using hidden markov models: A script-independent approach. 4768:231–250, 2008.

[5] W. Niblack. *An Introduction to Digital Image Processing*. Prentice Hall, Englewood Cliffs, NJ, USA,

1986.

[6] N. Otsu. A threshold selection method from gray-level histograms. *Systems, Man and Cybernetics, IEEE Transactions on*, 9(1):62 –66, Jan. 1979.

[7] X. Peng, S. Setlur, V. Govindaraju, and R. Sitaram. Binarization of camera-captured document using a MAP approach. In *SPIE Proc. XVIII Document Recognition and Retrieval*, volume 7874, pages 1–8, January 2011.

[8] J. Sauvola and M. Pietikäinen. Adaptive document image binarization. *Pattern Recognition*, 33:225–236, 2000.