International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), 4 Issue-5, November 2015

Printed Text and Handwriting Identification in Noisy Document Images

Mohammed Jassim Mohammed Jassim

Abstract- In this project, we address the problem of segmenting and identifying text in noisy document images. The identification of machine printed text and handwriting is important because: (1) recognition techniques available for machine printed text and handwriting are significantly different (2) Handwriting in a document indicates corrections, additions that should be treated differently from the main content. Instead of using simple noise filtering techniques, our approach treat noise itself is treated as a class. Thus our approach becomes a three-class identification problem (Machine Printed Text, Handwriting and Noise). After performing text identification, post processing is performed to refine classification accuracy.

Index Terms—Document Analysis, Printed Text, Segmentation, Markov Random Field, Hidden Markov Model.

I. INTRODUCTION

Documents are the written or printed-paper that contains layers such as letterhead, content, signature, annotations and noise. Document analysis tries to segment a document into layers with different physical and semantic properties. After decades of research, automatic document analysis has advanced to a point where text segmentation and recognition can be carried out in well-constrained documents. However, the performance degrades quickly when a small amount of noise is introduced. For example, a typical bottom up page segmentation method, first extract the connected components [1] and then group the connected components into text and zones if they are spatially close to each other. A classification process is then used to identify zone types (text, tables, images, etc.). These algorithms work well on clean documents where zones with different properties can be easily separated. However, they often fail on noisy documents. Previous work related to this problem has focused on text identification with the assumption that text regions are already segmented. The text identification is performed at the line, word or character level. Most of the researchers performed text identification at the line level [2, 3]. At the line level, the printed text lines are typically arranged regularly, while the handwritten text lines are irregular. Srihari et al. implemented a text line based approach and achieved the identification accuracy of 95% [2]. One advantage of the approach is it can be applied in different languages like Chinese, English etc with little or no modification. Although character level segmentation provides a very little information, Kuhnke proposed a neural network-based approach with straightness and symmetry as features, and achieved an identification accuracy of 96.8% and 78.5% [4]. Zheng used run-length histograms as features to identify handwritten and printed Chinese characters [5]. Based on the run-length histogram features, Zheng achieved the identification accuracy of 98%. However the same approach cannot be applied for Latin character identification. Some of the researchers performed text identification at the word level [6]. Guo et al. proposed an approach based on the vertical projection profile of the word [6]. They used a Hidden Markov Model (HMM) as the classifier and achieved the identification accuracy of 97.2%. In practice, however, handwritten annotations are often mixed with printed text.

In this paper, we perform word level segmentation using Bottom up approach, and then perform text identification. After performing text segmentation and identification, some handwritten characters are misclassified as noise and some noise are misclassified as handwriting. The misclassification cannot be avoided easily because the block provides only less information. So to avoid such misclassifications, Markov Random Field (MRF) based post processing is used which adds contextual information for each block and then refine classification. Contextual information is very helpful in improving classification accuracy. The idea is to model the statistical dependency of neighboring components. The Markov Random Field (MRF) is used to model the dependency of neighboring connected components. As post processing, MRFs can further improve classification accuracy.

The documents we are processing are extremely noisy with machine printed text, handwriting, and noise mixed together. We first extract the connected components and merge them at the word level based on spatial proximity. We then extract several categories of features and use trained Fisher classifiers to classify each word into machine printed text, handwriting, or noise. Finally, contextual information is incorporated into MRF models to refine the classification results further. The rest of the paper is organized as follows, section 1 describes the algorithms used, section 2 describes the steps involved in our project, section 3 compares our project with the existing system, and finally conclusion.

II. MATERIALS AND METHODOLOGY

In this project, 3 algorithms are used which are described as follows.

(1) Bottom up Approach

For performing word level segmentation, we use "Bottom up Approach" which works as follows.

- First extract the connected components.
- Estimate the average character size using component heights.

Revised Manuscript Received on 30 November 2015.
***** Correspondence Author
Dr. Mohamed Jassim Mohammed Jassim, Assistant Professor, College of Science and Information Technology, Iraq.

- Group the neighboring connected components into words if they spatially close to each other in horizontal direction. While merging two connected component into words, we enforce the following rules.

Two components C_1 and C_2 are merged only when they satisfy $\max(h_1, h_2) < 2 \times \min(h_1, h_2)$, where h_1 are h_2 are the heights of C_1 and C_2.

(2) Fisher Classifiers

It classifies each word as Machine printed text, Handwriting and noise. It is easier to train, faster for classification, needs only few training samples, and does not suffer from over training problems. It works as follows,

For a feature vector $\underline{X}$, the Fisher classifier projects $\underline{X}$ onto one dimension Y in direction $\underline{W}$. $Y = \underline{W}^T \underline{X}$

Let Y_1 and Y_2 be the projections of two classes and let $E[Y_1]$ and $E[Y_2]$ be the means of Y1 and Y2, respectively. Suppose $E[Y_1] > E[Y_2]$, then the decision can be made as

$$C(\underline{X}) = \begin{cases} class1 & if \quad Y > \left(E[Y_1] + E[Y_2]\right)/2 \\ class2 & Otherwise \end{cases}$$

(3) Markov Random Fields

Markov Random Field-based (MRF) based post processing is used to refine the classification accuracy. After performing classification, most of the blocks are correctly classified, but some of the blocks are misclassified in an overlapping feature space. It (**MRF**) improves classification accuracy by incorporating contextual information for each block. MRF is totally determined by clique and clique potential.

- Clique definition (Cp) for printed text defines the left and right neighbor of each block. Similarly for Clique definition for noise (C_n), except that it defines four neighbors for each block.
- Clique potential $V_p(c)$ and $V_n(c)$ is the energy associated with the clique. We assign high energy for the undesirable configuration and low energy for preferred configuration.

III. WORKING

The documents used, as an example in our project is extremely noisy with machine printed text, handwriting and noise mixed together. Our approach consists of four major steps, which are as follows,

- First we have to extract connected components, and then group the neighboring connected components into words based on the geometric size and proximity.
- After performing word level segmentation, we have to extract several sets of features like Run-length histogram features, Crossing count histogram features which shows the difference in stroke length between Machine printed text and handwriting. Texture features like bilevel co-occurrence features and Bilevel 2 × 2 gram features, and Structural features are extracted.
- After extracting several sets of features, classification starts. Here Fisher classifiers are used to classify each word as machine printed text, handwriting and noise. After performing classification, some misclassifications occur (i.e.) some noise blocks are misclassified as handwriting.
- The final step is Markov Random Field (MRF) based post processing for improving classification accuracy. It improves classification accuracy by incorporating contextual information for each block. The main idea is to model the dependency between neighboring components, which is performed by MRF and the classification is refined.

IV. COMPARISON WITH EXISTING SYSTEMS

Previous work related to this problem, filters noise by their size. It is not robust when the document is extremely noisy. But our approach on text segmentation and identification is robust enough in extremely noisy documents because, we extract more features for classification of machine printed text/handwriting and noise well. Also the previous work is a two-class identification problem (machine printed text and handwriting). But our approach is a three-class identification problem (machine printed text, handwriting and noise).

Also the existing system performs only text segmentation and classification. So some misclassifications (i.e. noise blocks are misclassified as handwriting and handwriting are misclassified as noise blocks) occur with the existing system. The classification accuracy is not good with the existing system. But our approach refines classification accuracy by adding contextual information for each block. Markov Random Field (MRF) based post processing provides contextual information for each block.

V. CONCLUSIONS

In this paper, we have presented an approach for segmenting and identifying text from extremely noisy document images. Instead of using simple filtering rules, we treat noise as a distinct class and use statistical classification techniques to classify each block into machine printed text, handwriting, and noise. We then use Markov Random Fields to incorporate contextual information for post processing. Our method is general enough to be extended to documents in other languages, with little or no modification. The technique presented in this paper can also be used for image enhancement to improve page segmentation accuracy of noisy documents. Currently, our clique potential definition considers only the labels of each block inside the clique, which may lose useful information. It is possible that one of the blocks is erroneously identified. Another potential improvement is to integrate high-level contextual information in addition to the local contextual information that we used. For example, the text line and zone segmentation results can be fed back to our classification module to refine the classification. This is the drawback in our approach.

REFERENCES

1. A.K. Jain and B. Yu, "Document Representation and Its Application to Page Decomposition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pp. 294-308, Mar. 1998.
2. S. N. Srihari, Y. C. Shim and V. Ramanaprasad. A system to read names and address on tax forms. Technical Report CEDAR-TR-94-2, CEDAR, SUNY, Buffalo, 1994.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

3. K. C. Fan, L. S. Wang and Y. T. Tu. Classification of machine-printed and handwritten texts using character block layout variance. Pattern Recognition, 31(9), pages 1275–1284, 1998.
4. K. Kuhnke, L. Summoning and Zs. M. Kovacs-V. A system for machine-written and handwritten Character distinction. In Proc. of the 3rd Inter. Conf. On Document Analysis & Recognition, Pages 811–814, 1995.
5. Y. Zheng, C. Liu and X. Ding. Single character type identification. In Proc. of SPIE Vol.4670, Document Recognition & Retrieval IX, pages 49–56, 2001.
6. J. K. Guo and M. Y. Ma. Separating handwritten material from machine printed text using hidden Markov models. In Proc. of the 6th Inter. Conf. On Document Analysis & Recognition, pages 439–443, 2001.
7. www.mathworks.com

Dr. Mohammed Jassim Mohammed Jassim, received is PhD in 2007,an Assistant Professor in the College of Science and Information Technology, Iraq. He published 9 Research Paper in International Journals, participated in 7 International Conferences and translated 12 books from Russian to Arabic.

FLOWCHART