Writer Identification For Handwritten Document Based On Structured Learning

Miss. Kiran Gole, Miss. Julekha Mulani, Mr. Govind Kumar, Prof. Deepali Gosavi
ISB & M School Of Technology Nande Village, Tal.Mulashi, Pune, Savitribai Phule Pune University
Kirangole7@gmail.com

Abstract: For OCR (optical character recognition) and indirectly to identify the original writer of the handwritten document, segmentation of handwritten document images into text-lines and words is an essential task. Since the features of handwritten document are irregular and is different depending on the person therefore it is considered a challenging problem. To address the problem, we formulating the problem of word segmentation as a binary quadratic assignment problem that considers pair wise correlations between the gaps in the text and also of individual gaps. Using the Structured SVM (Support Vector Machine) framework we estimate all the parameter to work the proposed method well regardless of different writing styles and written languages without user-defined parameters.

Keywords: Handwritten Document, word segmentation, SVM, binary quadratic assignment,text-lines.

I. Introduction

To understand the handwritten document and also to match the document the segmentation of document into the word and text-line is the important fact today. As compare to machine-printed document the handwritten document is more challenging due to-

(i) irregular spacing between words and

(ii) variations of writing styles depending on the person.

To identify and verify writer we developed effective technique which uses probability distribution function (PDFs) extracted from handwriting images to identify particular or exact writer.

To solve the problem, this paper proposes a scale invariant feature transform (SIFT) based method to extract the key based structural features from handwriting images. Main two features-

1.SO(SIFT Orientation)

2.SD (SIFT Descriptor)

A. Conventional Approaches for the Word Segmentation

For the word segmentation, document images are first segmented into text-lines .Then, the word segmentation algorithm (for a single text-line) is applied to individual text-lines. Given a single text-line, the conventional word segmentation algorithms consist of two steps: The first step is to extract candidates for inter-word gaps (word- separator) and the next step is to classify the candidates into intra/inter-word gaps. For the candidate generation, a given text-line is represented with a set of super-pixels (where a super-pixel usually corresponds to a letter or a group of letters) and their gaps are considered candidates to be classified. This is a binary classification problem that assigns a label 0,1, where 0 means that the gap is an intra-word gap and 1 indicates it is an inter-word gap. For this classification, many algorithms have been developed: global/adaptive thresholding was used in, the unsupervised learning techniques such as clustering and Gaussian Mixture Model (GMM) were adopted in and the scale space selection approach was employed in. Also, there have been researches using supervised-learning techniques such as neural networks . However, they only considered the local properties of individual gaps (without the considerations on correlations between the gaps).

B. Our approach

Although the characteristics of inter-word gaps are changing across (and even in) documents, there are strong correlations (e.g., scale) between them in a text-line. However, it has been difficult to exploit these correlations in the conventional approaches, where the classification is made independently based on the properties of each gap. In order to alleviate these problems, we develop a novel framework that considers these correlations as well as local observations (i.e., the properties of each gap). To be precise, we formulate the word segmentation as an optimization problem that maximizes the similarity between inter-word gaps and the dissimilarity between inter-word and intra-word gaps, in addition to the likelihoods. Since this problem is a binary classification problem and the singleton and pair wise terms are only considered, it can be formulated as a binary quadratic problem, which can be efficiently solved with the Mixed-Integer Quadratic Programming (MIQP) solvers.

Also, we estimate all parameters by adopting the structured learning framework , so that the proposed method can deal with a variety of inputs without user-defined parameters. Therefore, we believe that the main contributions of our project :

(i) The novel formulation of the word segmentation problem into a binary quadratic problem

(ii) Improved performances on challenging datasets.

II. Material and Methodology

We are solving the problem of word segmentation using SIFT algorithm. We are dividing the solution in three parts these are:

1.Enrollment:

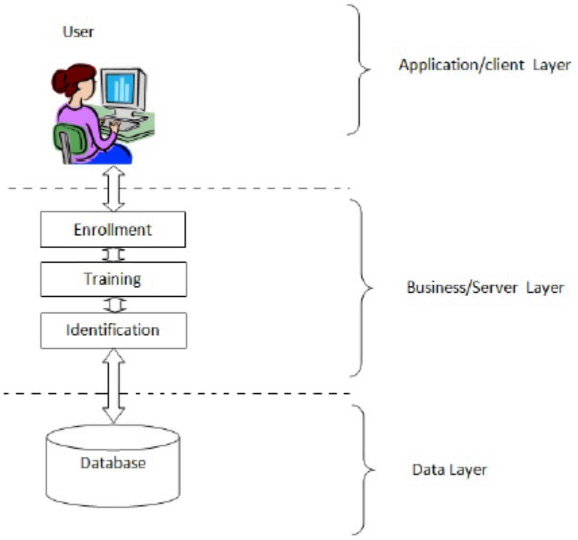
This phase register the user when user will access the system for first time, the user will get the password and the unique user id, through which user will access the system.

2.Training:

The user will access the system using provided id and

doi : 10.17950/ijer/v5s1/113

password, also user give input as the image of handwritten document to identify writer of that document.



3.Identify:

In this phase the algorithm will be applied on image to extract the features of image ,also this phase gives name of the writer of the document.

Algorithms and methods use to solve the problem:

1. Text-Line Segmentation and Super-Pixel Representation This algorithm is use to break the text-line in the Image into segments. The Segments(group of pixel)are also represented using this algorithm.

2. Structured learning for word segmentation The feature vector is represented using this method to achieve structure learning, the feature vector is nothing but the features of the text in the image

3. Algorithm for Cutting Plane:

As the solution will having of large in size so to improve the efficiency cutting plane algorithm is used

4. SIFT algorithm: The main algorithm that use to solve the problem is the SIFT Algorithm .The feature extraction and the comparison of document to identify the writer of document.

III. Results

In today's world segmentation of document images into text-lines and words is an important step for the document understanding .But the segmentation of handwritten documents is still considered a challenging problem due to irregular spacing between words and variations of writing styles depending on the person. Using the text-line segmentation algorithms have been armature to some extent, however, there is still need of improvements in the case of word segmentation methods.To Identify the writer of handwritten document the segmentation of text-line is the important task ,so we are adopting the structured learning methods and the MIQP method(Multiple Integer Quadratic Problem)to solve the problem of word segmentation and identifying the correct writer of document

IV. Conclusion

In this paper , For handwritten document images we have proposed a SIFT Algorithm and Word segmentation algorithm. By using the SIFT algorithm the features of text get extracted. We consider the segmentation problem as a binary quadratic programming and estimate the parameters using structured learning. Due to the proposed formulation, we focus on pair-wise similarities between word-separator and also unary properties in the word segmentation. By using the Structured SVM, all parameters are estimated

Acknowledgement

Authors are cordially giving thanks to the researchers of different model for Text Extraction from Handwritten Document. All other who have tried hard to make their work easy to accomplish.

References

i. *"Word segmentation method for handwritten document based on structure learning" Jewoong Ryu, Hyung Il Koo, Member, IEEE, and Nam Ik Cho, Senior Member, IEEE,8 aug -2015*

ii. *"Task-Speci_c Image Partitioning" Sungwoong Kim*, Member, IEEE, Sebastian Nowozin, Pushmeet Kohli, and Chang D. Yoo, Senior Member, IEEE*

iii. *"Language Independent Text-Line Extraction Algorithm for Handwritten Documents" Nikita Vijay Borse, Prof. Imran R. Shaikh Department of CSE, S.N.D.C.O.E.R.C Yeola, Dist-Nashik,Savitribai Phule -University of Pune.*

iv. *"A New Segmentation Algorithm for Online HandwrittenWord Recognition in Persian Script"Sara Izadi, Mehdi Haji, Ching Y. Suen Centrefor Pattern Recognition and Machine Intelligence, 1455 de Maisonneuve Blvd. West, Montral, Qubec, Canada, H3G1M8 fs izadin, mhaji, sueng@cs.concordia.ca.*

v. *5. "Handwritten Text Segmentation using Average Longest Path Algo-*

vi. *rithm" Dhaval Salvi, Jun Zhou, Jarrell Waggoner, and Song Wang Department of Computer Science and Engineering University of South Carolina, Columbia, SC 29208, USA.*

vii. *"Handwritten document imagesegmentation into textlinesandwords" Vassilis Papavassilioua,b,?, ThemosStafylakisa,b, VassilisKatsourosa,GeorgeCarayannisa.*

viii. *Text-line extraction in handwritten Chinese documents based on an energy minimization framework, H. I. Koo and N. I. Cho,IEEE Trans. Image Process., vol. 21, no. 3, pp. 116975, Mar. 2012.*

ix. *Handwritten document image segmentation into text lines and words, V. Papavassiliou, T. Stafylakis, V. Katsouros, and G. Carayannis,Patt. Recognit., vol. 43, no. 1, pp. 369377, Jan. 2010*

x. *Mixed-integer quadratic programming, R. Lazimy,Math. Program. vol.22, no. 1, pp. 33234*