2

State-of-the-Art : Printed vs Handwritten

In our work, we primarily addressed the problems of annotations extraction and writer identification. This chapter presents a survey of the past work related to the classification of handwritten and printed text. In the next chapter, we present a review of the existing methods for the writer identification task. The organization of this chapter is as follows. Section 2.1 examines the distinctiveness in the shape and statistics of the printed and handwritten content, which has motivated us to explore the separation of handwritten and machine printed text. In the subsequent sections the state-of-the-art methods are described in terms of their characteristics, constraints and the different datasets they use. Post processing techniques, which help in improving the accuracy, are also examined. Section 2.2 reviews the classification at character level. In character-level segmentation the whole document is broken down into characters and each character is assigned a label as printed or handwritten. Section 2.3 reviews the classification of larger text blocks as printed or handwritten. Section 2.4 reviews the classification of words as handwritten or printed. Correct segmentation of words is sometimes difficult, hence textual connected components are extracted. Section 2.5 describes the classification of pseudo words or connected components. Section 2.6 gives an overview of pixel-level classification schemes to separate annotations overlayed on the printed text. Section 2.7 describes the methods to classify the text-lines as handwritten or printed. Section 2.8 gives an overview of the post processing methods to improve the rate of classification. Section 2.9 concludes the chapter with a discussion on the problems that have remained largely unattended.

2.1 INTRODUCTION

Automatic segmentation and layout analysis of documents can be used for applications like interpretation and machine translation of technical documents, search and information retrieval, and mobile-based document reading order determination. Apart from its utility in OCR processing, layout segmentation is also widely applied for document compression, document summarization, multilingual document analysis, and ink beautification.

Layout segmentation can be studied as geometric layout segmentation and logical layout segmentation [Haralick, 1994]. The former suggest specifying geometry of the homogeneous regions and their relationship in a document while the latter suggests semantic labeling of the blocks of the page [Haralick, 1994]. It is challenging to segment text and non-text regions; and it is more difficult to segment the text as printed and handwritten. From the comprehensive survey stated in [Govindan and Shivaprasad, 1990; Impedovo *et al.*, 1991; Kuhnke *et al.*, 1995; Wong *et al.*, 1982], it can be understood that machine-printed and hand-written recognition schemes are quite different from each other. Table 2.1 illustrates the visual differences between machine printed and handwritten text. So, if a document consists of both machine-printed and hand-written portions, they should be separated before feeding them to the respective OCR systems [Pal and Chaudhuri, 1999, 2001].

Storage of enormous electronic data from scanning books and manuscripts is a concern, which is eased by data compression techniques. For efficient compression, the document image

Table 2.1 : Visual Differences Between Machine Printed and Handwritten Text.

Printed Text	Handwritten Text
Pixel intensity values are almost similar for a whole word.	Pixel intensity values are varies because of varying hand pressure while writing.
Envelope straightness persist throughout the word/text	The envelopes have large Crests and troughs.
Character shape is unique in each font type.	The text shape depends on each individual person.
The size of characters (height and width) is constant.	The size of characters (height and width), is not constant.
Inter character word spacing is constant.	Inter character/word spacing always varying.
Number of black pixels to represent a character is constant for a particular font type and font size.	Due to varying hand pressure the number of black pixels to represent a character is not constant.
The variance in intensity values are less compare to handwritten.	The variance in intensity values are more compare to machine-printed.
The stroke width is constant for a particular font type and font size.	Stroke width is never constant.

is segmented into its subsequent sub-images comprising text, painted pictures, photographs etc. [Xu and Bao, 2009]. Different parts of a page need different schemes for compression [Garain *et al.*, 2003] and hence geometric segmentation turn out to be need of the hour for such domain. Moreover, essentially in any practical handwriting recognition system geometric segmentation is desired as well. This is because handwriting is unconstrained and depends on writers. From Postal Addresses Interpretation Systems to Bank-Cheque Reader Machines; from Forensic Document Examination [Srihari and Leedham, 2003] to Automated Forms Processing, anchor detection [Likforman *et al.*, 2006] segmentation of the hand written and printed text is looked-for as a preprocessing step. Sometimes we annotate the books while reading and add explanatory notes to it. Those annotations are valuable information to further summarize the document. Likewise sentiment analysis can be achieved by analyzing the handwritten annotations over the printed text.

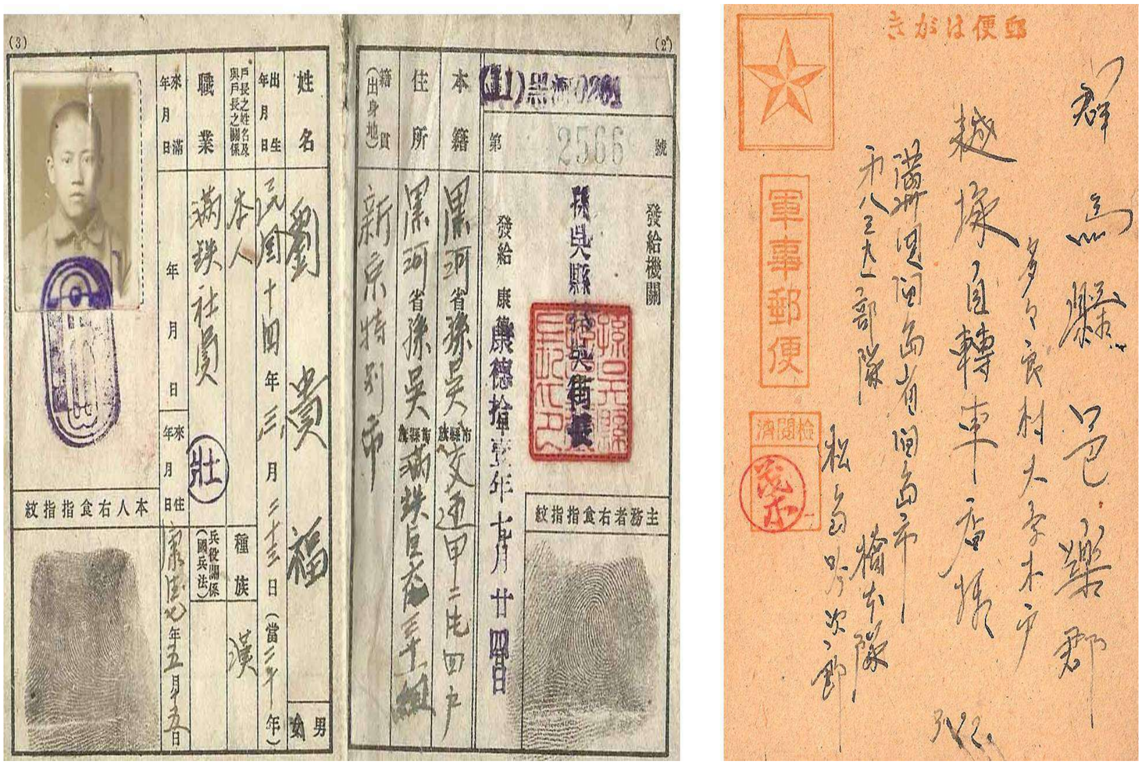
2.2 CLASSIFICATION AT CHARACTER LEVEL

While at the character level less information is available; humans can still identify the handwritten and printed characters easily. This keen observant power of human brain inspires researchers to pursue classification at the character level. Figure 2.1 demonstrates examples of such documents where the text is broken to characters and encourages segmentation of handwritten and printed content at character level.

Table 2.2 summarizes the work for character level classification of Printed and Handwritten text in chronological order.

Imade *et al.* [1993] separated Chinese characters using histograms of gradient vector directions and luminance features. The image is divided into blocks, with an assumption of each block is nearly encompassing a character. The features are fed into a NN to classify the blocks as printed or handwritten. Their scheme produces an accuracy of 48% for printed and 45.5% for handwritten character.

Kuhnke *et al.* [1995] studied the structure of Roman characters and found straightness property in their contours. Hough transform was applied to measure the straightness among



(3)
姓名 劉賞福
男 女
出生年月日 民国十四年三月二十三日（當二十一年）
與戸長之關係及戸長之姓名 本人
職業 満鉄社員
種族 漢
兵役關係（國兵法）
來滿年月日 年 月 日
來住年月日 康徳七年五月十五日
壯
本人右食指指紋

(2)
（1）黑
第 2566 號
發給機關 孫呉縣
發給 康徳拾壹年十月廿四日
本籍 黑河省孫呉 市縣旗 支道甲二屯四戸
住所 黑河省孫呉 市縣旗 満鉄旦宅第一組
籍貫（出身地） 新京特別市
主務者右食指指紋

郵便はがき
軍事郵便
検閲済
群馬縣邑樂郡
多々良村大字木戸
越塚自轉車店様
満洲国間島省間島市
九八三九部隊
橋本隊
松島好次郎

Figure 2.1 : Examples of such documents [warrelics.eu, 2010; Japanese Military Postcards, 1931] in real environment where the text is broken to characters and encourages segmentation of handwritten and printed content at character level.

the Roman characters of NIST Special Database. A feed forward network was adopted as the recognition scheme and it achieved an accuracy of 78.5% on the test set. However, the recognition performance was poor for italics characters.

Character-based procedures involve segmentation as a preprocessing step, which may be error-prone. In view of this, [Koyama *et al.*, 2008] proposed a segmentation free method that locally captures the fluctuations caused by handwriting. A window nearly equal to the character size selects local regions from the image. The spatial information within a region is transformed into frequency domain for features extraction. Evidently, machine-printed characters have ordered patterns of line segments while hand-written characters have unevenness of line segments. This method is called spectrum-domain local fluctuation detection (SDLFD) method. Finally a multi layer perceptron (MLP) was used to learn the features and it gave an accuracy of 97% for Alphanumeric Hiragana, Katakana and Chinese characters. Their scheme is robust against variation in scanning resolution.

Hitherto, the classification of handwritten and printed text is restricted to mono-lingual documents. We now review the work that has addressed detection of handwritten characters in multiple languages. Song *et al.* [2011a] proposed a technique that can extract handwritten characters written in multiple languages from a printed document. First, the words are extracted by morphological connected component analysis. These connected components are then grouped together by choosing appropriate merging rules based on the language. Since the spatial proximity of the words can be quite distinct for different languages, different merging rules are applied for connected components of different languages. This made it necessary to first identify the language of connected components. Their work focused on three languages: Chinese, English and

Japanese, which can be categorized into two classes, using strokes-composed or letter-composed feature. Finally, words are segmented, and four kinds of features are extracted, including structure feature, run-length feature, cross-count feature and bi-level co-occurrence feature. To further improve results, feature optimization is performed by feature fusion algorithm, and Genetic algorithm is applied for type classification. Finally, a Markov Random Field model is utilized as a post-processing step to further correct the misclassification of word type by considering the document context. Their work reported a precision of 99.25% for English, 97.58% for Japanese, an 97.10% for mixed content.

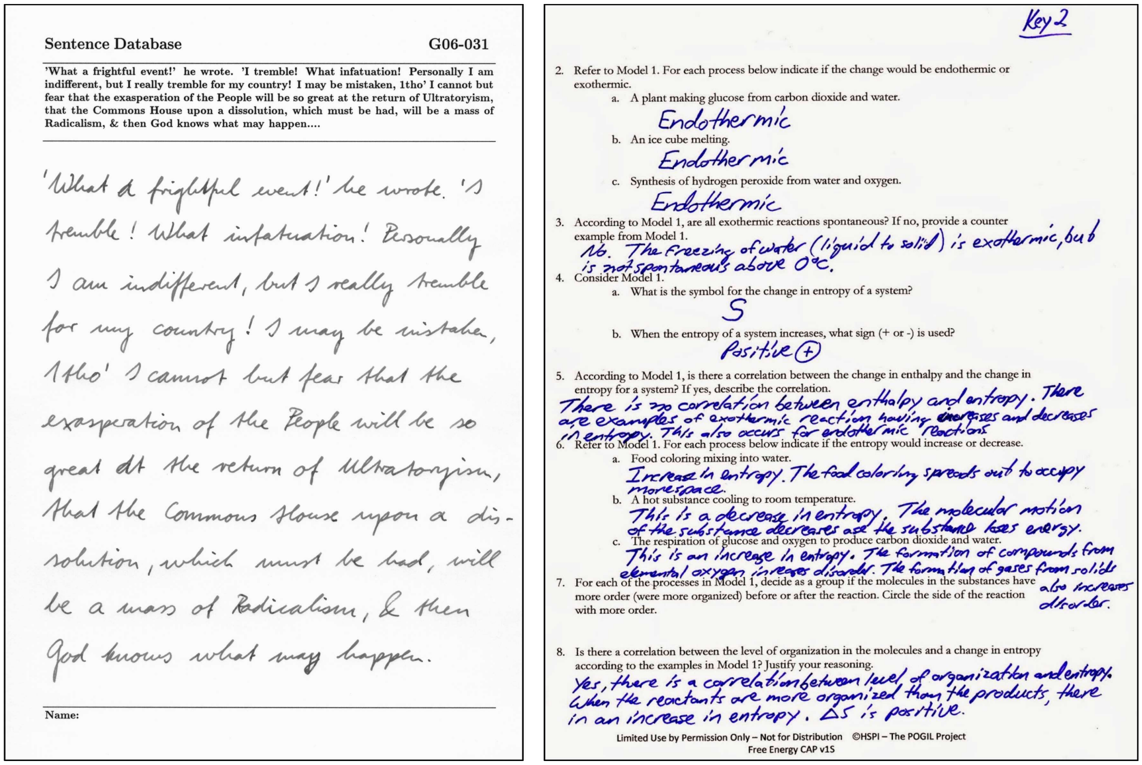
Table 2.2 : Overview of dataset, performance and methods applied for handwritten character classification from printed text.

System	Dataset	Segmentation/ Classification Scheme	Training Data	Testing Data	Performance	Language
Imade *et al.* [1993]	Kanji and Kana	NN	…	…	48%(P), 45.5%(H)	Kanji and Kana character
Kuhnke *et al.* [1995]	NIST Special Database	Feed-forward Neural Network	3652 characters	1068 characters	78.5%	Roman alphabet
Koyama *et al.* [2008]	ETL character databases	MLP	2000 characters	500 characters	78.5%	Alphanumeric, Hiragana, Katakana, Chinese character
Song *et al.* [2011a]	Self-created dataset, IAM and Tobacco dataset	Genetic Algorithm	…	…	99.42% (Chinese), 99.25% (English), 97.58% (Japanese), 97.10% (Mixed)	Chinese, English, Japanese character

2.3 CLASSIFICATION AT BLOCK LEVEL

Now-a-days the documents have become larger in size and therefore require a longer processing time. Rather than disintegrating a document into its atomic units like characters, words, text-lines and then classify as printed/handwritten, it would be fast if we do classification of large chunks/blocks. Figure 2.2 demonstrates examples of such documents in real environment where large chunks of single text persist and encourages segmentation of handwritten and printed content at block level. Table 2.3 summarizes the handwritten and printed text-block classification methods in chronological order.

Violante *et al.* [1995] describe a computationally efficient technique for discriminating between hand-written and printed text on mail. The image is cleaned and foreground pixels are separated. These pixels are glued together to a count of fixed size to form regions. From these regions, edge straightness and profile features are extracted and fed into a NN for label prediction.



Sentence Database G06-031

'What a frightful event!' he wrote. 'I tremble! What infatuation! Personally I am indifferent, but I really tremble for my country! I may be mistaken, 1tho' I cannot but fear that the exasperation of the People will be so great at the return of Ultratoryism, that the Commons House upon a dissolution, which must be had, will be a mass of Radicalism, & then God knows what may happen....

'What a frightful event!' he wrote. 'I
tremble! What infatuation! Personally
I am indifferent, but I really tremble
for my country! I may be mistaken,
1tho' I cannot but fear that the
exasperation of the People will be so
great at the return of Ultratoryism,
that the Commons House upon a dis-
solution, which must be had, will
be a mass of Radicalism, & then
God knows what may happen.

Name:

Key 2

2. Refer to Model 1. For each process below indicate if the change would be endothermic or exothermic.
   a. A plant making glucose from carbon dioxide and water.
      Endothermic
   b. An ice cube melting.
      Endothermic
   c. Synthesis of hydrogen peroxide from water and oxygen.
      Endothermic
3. According to Model 1, are all exothermic reactions spontaneous? If no, provide a counter example from Model 1.
   No. The freezing of water (liquid to solid) is exothermic, but is not spontaneous above 0°C.
4. Consider Model 1.
   a. What is the symbol for the change in entropy of a system?
      S
   b. When the entropy of a system increases, what sign (+ or -) is used?
      Positive (+)
5. According to Model 1, is there a correlation between the change in enthalpy and the change in entropy for a system? If yes, describe the correlation.
   There is no correlation between enthalpy and entropy. There are examples of exothermic reaction having increases and decreases in entropy. This also occurs for endothermic reactions
6. Refer to Model 1. For each process below indicate if the entropy would increase or decrease.
   a. Food coloring mixing into water.
      Increase in entropy. The food coloring spreads out to occupy more space.
   b. A hot substance cooling to room temperature.
      This is a decrease in entropy. The molecular motion of the substance decreases as the substance loses energy.
   c. The respiration of glucose and oxygen to produce carbon dioxide and water.
      This is an increase in entropy. The formation of compounds from elemental oxygen increases disorder. The formation of gases from solids also increases disorder.
7. For each of the processes in Model 1, decide as a group if the molecules in the substances have more order (were more organized) before or after the reaction. Circle the side of the reaction with more order.
8. Is there a correlation between the level of organization in the molecules and a change in entropy according to the examples in Model 1? Justify your reasoning.
   Yes, there is a correlation between level of organization and entropy. When the reactants are more organized than the products, there in an increase in entropy. ΔS is positive.

Limited Use by Permission Only – Not for Distribution ©HSPI – The POGIL Project
Free Energy CAP v1S

Figure 2.2 : Examples of such documents in real environment where large chunks of text persists and encourages segmentation of handwritten and printed content at Block Level.

An accuracy of 95% was achieved for discrimination of handwritten and printed text.

Following the application of mail-address processing [Wolf *et al.*, 1997] also extracted handwritten address blocks from a set of mail-pieces. Fixed sized blocks were segmented from the mail envelops and mean gradient magnitude was computed for each block. This magnitude on comparison with a threshold classifies homogeneous printed blocks from inhomogeneous handwritten blocks. This heuristic approach when applied on 2000 envelopes, produced an accuracy of 91.3%.

Kumar *et al.* [2011] proposed a codebook method to extract handwritten and printed text zones from noisy Arabic document images with an accuracy of 91%. Vornoi algorithm was applied to segment zones from a document. The shape characteristic of each zone is exploited using a novel edge based feature called Triple-Adjacent-Segment (TAS). Using this shape feature individual codebooks are constructed from a set of handwritten and printed text documents. The final descriptor is constructed by concatenating the individual normalized histogram from each codebook. An SVM classifier is trained to classify the respective zones. The TAS features are invariant to translation, scale and rotation of text. The method is robust to the background noise present in the image.

Zagoris *et al.* [2014] addressed the problem of handwritten and machine-printed text separation using the Bag of Visual Words (BoVW) model. Their work can be divided into three stages. In the first stage the graphical content is removed from the image and text blocks are extracted using adaptive RLSA. In the second stage the block descriptors are computed. An optimal codebook using a Self-Growing and Self-Organized Neural Gas (SGONG) network is created with SIFT descriptors of handwritten and printed blocks. The third stage is a combination of two 1 vs rest SVM classifiers. The output of each SVM is analyzed and the final classification of each block descriptor is made as printed, handwritten or noise. The experiments are performed on PRImA NHM dataset and IAM dataset. The method produced an accuracy of 84.2% for PRImA dataset and 98.9% for IAM dataset. In terms of classification, the algorithm outperforms multi-class SVM and Random Forest approaches. Also, in terms of feature separability for each class, the approach is less vulnerable to segmentation failures compared to the baseline Gabor features.

Emambakhsh *et al.* [2016] presented a template matching approach for the discrimination of handwritten and machine-printed text blocks. The method does not require training and is robust against occlusions and noise. The document image is cleaned and text blocks are segmented. A large gallery containing various characters written using different machine font attributes is generated. This flexible sized gallery is used to align and match characters with the segmented regions in a parallel fashion, using normalized cross-correlation. If the output of the matching score is higher than a pre-defined threshold, the region underlying the mask in which highest matching score is achieved is labeled as machine-printed and excluded from the image. If for all of the gallery samples the matching scores are lower than the pre-defined threshold, the text block is labeled as handwritten. The experimental results over PRImA-NHM dataset show 84.0% classification rate in classifying cluttered, occluded and noisy samples.

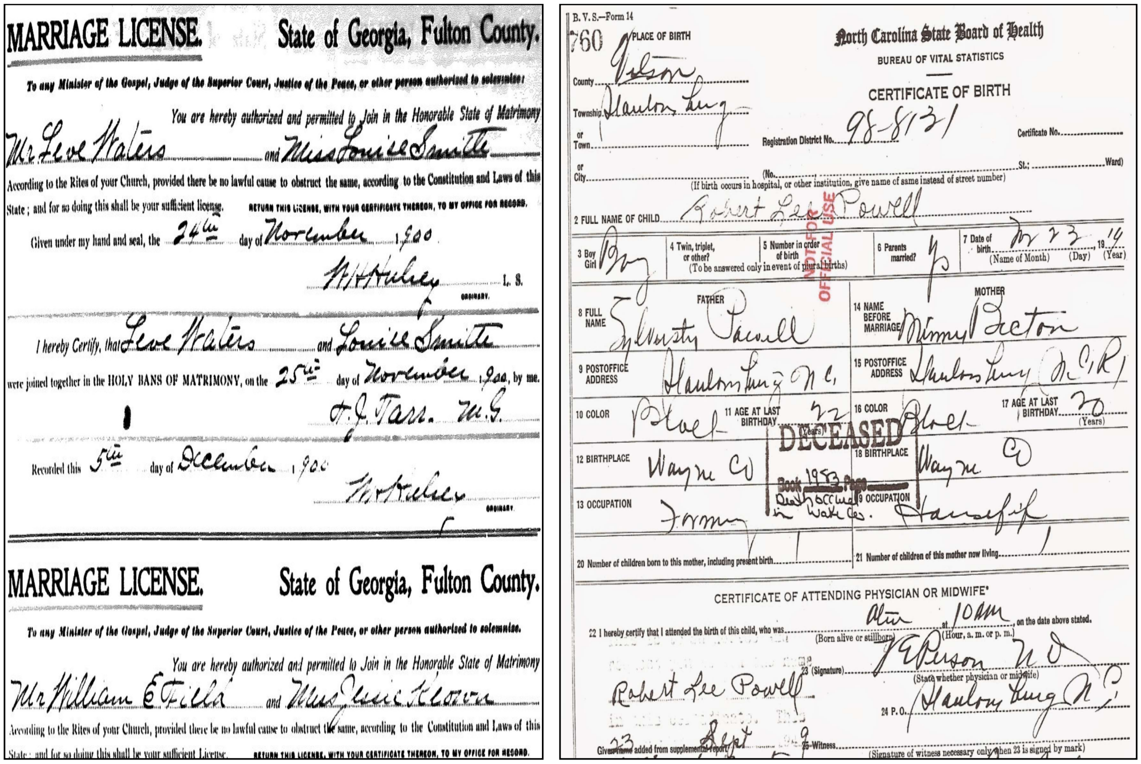
Table 2.3 : Overview of Dataset, Performance and Methods applied for Classification of Handwritten Text Blocks from Printed Text Blocks.

System	**Dataset**	**Segmentation/ Classification Scheme**	**Training Data**	**Testing Data**	**Performance**	**Language**
Violante *et al.* [1995]	Mail envelopes	NN	…	…	95%	English
Wolf *et al.* [1997]	Mail-pieces	Heuristic Approach	…	2000 envelops	91.3%	English
Kumar *et al.* [2011]	10,946 Documents	Codebook method with SVM	732	625 Documents	91%	Arabic
Zagoris *et al.* [2014]	PRImA and IAM Dataset	BoV with SVM	36 Documents	200 Documents	84.2% PRImA, 98.9% IAM accuracy	English
Emambakhsh *et al.* [2016]	PRImA	Template Matching	…	100(H), 415(P) segments,	84%	English

2.4 CLASSIFICATION AT WORD LEVEL

Documents containing mixed types of text (printed and handwritten) are increasingly present in business and academic environments. They result frequently from annotating printed documents such as bills, administrative forms, birth-certificates, letters, etc.

Often text discrimination as printed and handwritten is done at the word level because the the best classification is achieved at word level [da Silva *et al.*, 2009]. Such type of segmentation level is applicable where the text lines are composed of both printed and handwritten text. Figure 2.3 demonstrates examples of such documents in real environment where mixed types of text persists and encourages segmentation at word level. Table 2.4 summarizes the word-based handwritten and printed text classification methods in chronological order.



MARRIAGE LICENSE. State of Georgia, Fulton County.

To any Minister of the Gospel, Judge of the Superior Court, Justice of the Peace, or other person authorized to solemnize:

You are hereby authorized and permitted to Join in the Honorable State of Matrimony

Mr Leve Waters and Miss Louise Smith

According to the Rites of your Church, provided there be no lawful cause to obstruct the same, according to the Constitution and Laws of this State; and for so doing this shall be your sufficient license. RETURN THIS LICENSE, WITH YOUR CERTIFICATE THEREON, TO MY OFFICE FOR RECORD.

Given under my hand and seal, the 24th day of November 1900

W H Hulsey L. S.
ORDINARY.

I hereby Certify, that Leve Waters and Louise Smith were joined together in the HOLY BANS OF MATRIMONY, on the 25th day of November 1900, by me.

T. J. Tarr. M.G.

Recorded this 5th day of December 1900

W H Hulsey
ORDINARY.

MARRIAGE LICENSE. State of Georgia, Fulton County.

To any Minister of the Gospel, Judge of the Superior Court, Justice of the Peace, or other person authorized to solemnize.

You are hereby authorized and permitted to Join in the Honorable State of Matrimony

Mr William E Fields and Miss Jessie Gloven

According to the Rites of your Church, provided there be no lawful cause to obstruct the same, according to the Constitution and Laws of this State; and for so doing this shall be your sufficient License. RETURN THIS LICENSE, WITH YOUR CERTIFICATE THEREON, TO MY OFFICE FOR RECORD.

B. V. S.—Form 14

760 PLACE OF BIRTH

North Carolina State Board of Health
BUREAU OF VITAL STATISTICS
CERTIFICATE OF BIRTH

County Wilson
Township Stantonsburg
or Town
Registration District No. 98-8131
Certificate No.
or City (No. St.; Ward)
(If birth occurs in hospital, or other institution, give name of same instead of street number)

2 FULL NAME OF CHILD Robert Lee Powell

NOT FOR OFFICIAL USE

3 Boy Girl Boy | 4 Twin, triplet, or other? | 5 Number in order of birth (To be answered only in event of plural births) | 6 Parents married? yes | 7 Date of birth Nov 23, 1914 (Name of Month) (Day) (Year)

FATHER
8 FULL NAME Sylvester Powell
9 POSTOFFICE ADDRESS Stantonsburg N C,
10 COLOR Black
11 AGE AT LAST BIRTHDAY 22 (Years)
12 BIRTHPLACE Wayne Co
13 OCCUPATION Farmer

MOTHER
14 NAME BEFORE MARRIAGE Minnie Barton
15 POSTOFFICE ADDRESS Stantonsburg N C R 1
16 COLOR Black
17 AGE AT LAST BIRTHDAY 20 (Years)
18 BIRTHPLACE Wayne Co
19 OCCUPATION Housewife

DECEASED
Book 1953 Page
Death occurred in Wake Co.

20 Number of children born to this mother, including present birth 1
21 Number of children of this mother now living 1

CERTIFICATE OF ATTENDING PHYSICIAN OR MIDWIFE*

22 I hereby certify that I attended the birth of this child, who was Alive (Born alive or stillborn) at 10 AM (Hour, a. m. or p. m.), on the date above stated.

23 (Signature) T E Person M D
(State whether physician or midwife)

Robert Lee Powell

24 P. O. Stantonsburg N C

Given name added from supplemental report 23 Sept 1949
25 Witness
(Signature of witness necessary only when 23 is signed by mark)

Figure 2.3 : Examples of such documents [Genealogy Photos, 2010; Birth Certificate, 2018] in real environment where mixed types of text persists and encourages segmentation of handwritten and printed content at word/connected component Level.

Guo and Ma [2001] applied HMM as a classification scheme to classify pre-segmented words. Since HMM is associated with sequential learning, it requires constructing an observation sequence for the word. This was achieved by segmenting a word into individual letters represented by bounding boxes. The letter in each word can form the observational sequence. The observation sequence was obtained by concatenating projection profiles from individual letters that are within the word. Post processing based on neighborhood scheme further improves the results. They achieved a precision of 92.86% and a recall of 72.19% for 187 handwritten words. However, the method identified overlapping words as handwritten words, and also, the method is difficult to adapt to the structure of languages like Chinese and Japanese.

Presence of noise usually produces erroneous results and hence needs to be addressed. Zheng *et al.* [2004] trained Fisher classifiers to classify each word into machine printed text, handwriting, and noise. Morphological operations were used to find CC which are merged to form complete words by applying spatial proximity. For each word, properties such as stroke orientation, complexity and length were extracted along with texture and structural properties. Finally, contextual information was incorporated into MRF models to further refine the classification result. Evaluation was done on Tobacco Industry dataset consisted of 318 annotated documents, and the method achieved an overall accuracy of 97.3%. Although the system has acceptable accuracy, yet it classifies the noise and other non-text regions as handwritten words.

Another work that makes use of only texture features to develop a font independent scheme was reported in [Farooq *et al.*, 2006]. They extracted a 12 dimensional feature vector by applying 12

Gabor filters at different scales and orientations to classify words as printed or handwritten. These features are fed into a probabilistic neural network to produce classification precision of 94.62% on Arabic script.

A rule-based classification scheme was developed by da Silva *et al.* [2009] to classify a word as printed or handwritten. They applied the WEKA rule-based classification system to segment handwritten words from IAM datset. They mined structural features by WEKA tool during the training phase of the system. They also experimented on their own dataset. Their work reports an overall precision of 97% on printed and recall of 98% in handwritten words. However, their dataset is free from logos, figures, tables, graphs or another type of element.

Zemouri and Chibani [2011] exploited the structural properties of printed or handwritten characters in word and suggested to use the Radon transform as a discriminating feature. They observed that the shape of printed characters is more or less stable within a text word while the distribution of the shape of handwritten characters is quite diverse. Radon features are concatenated with statistical features trained an SVM classifier for a robust separation of the two classes. They reported an accuracy of 98.08% on the IAM dataset.

In real world scenario documents are generally composed of noise and graphical contents. The work by Belaïd *et al.* [2013] aimed to separate the printed and handwritten text on documents with noise and graphics. They used multiclass SVM for classification and applied *k*-d trees for context labeling as a post processing scheme. Their work achieved an accuracy 90% on a smaller dataset. Though they classified words but their evaluation measure was based on pixels hence all test documents got perfectly labeled at pixel level.

Awal *et al.* [2014] addressed the problem of printed and handwritten word separation in real noisy documents. Connected component analysis was used to break down the document into pseudo lines and then to pseudo words. Their work primarily aimed to introduce and compare post processing schemes to classify the pseudo words as printed, handwritten or noise. To extract pseudo lines and pseudo words a proximity string segmentation algorithm was used. A vector of 137 features was extracted from each pseudo-word and a multi-class SVM with Gaussian kernel was used to assign an appropriate label. The results are improved by contextual relabeling which results in an accuracy of 97.3% and 99.5% for handwritten and printed word respectively. However, there is further scope to improve the performance for documents comprising overlapping text and graphical contents.

Segmentation of handwritten text in not limited only to a single script document. Echi and Saidani [2014] could segment Arabic and Latin handwritten words when trained together by Bayes (AODEsr) classifier. They also compared performances of five classifiers: Bayes (AODEsr), k-NN, Decision Tree, SVM, MLP. Among them, Bayes (AODEsr) is rated best with an average precision and recall rates for identification as 98.72% using a set of 58 features, which included word-based profile features, structural features, connected component profile features, run-length and co-occurrence features.

Ricquebourg *et al.* [2014] applied boosting for discriminating handwritten/printed words and compared it with an SVM classifier . They used high dimensional feature set for classification. Features included width, height, surface, pixel-value average, center of inertial coordinates, moments of inertia, Zernike moments and histograms of 8-contour directions using Freeman chain code representation. Their work reported an overall precision of 95.28% using the Bonsai Boosting method which is comparable to SVM, which reports a precision of 95.90%.

Malakar *et al.* [2013] applied decision tree with feature ranking to classify the printed and handwritten words. They used six gray-level statistical features and extracted a feature vector of six dimensions from each word. They achieved an overall precision of 96.8%. However, noisy words were most of the time misclassified.

Table 2.4 : Overview of Dataset, Performance and Methods applied for Handwritten Words Classification from Printed Text.

System	Dataset	Segmentation/ Classification Scheme	Training Data	Testing Data	Performance	Language
Guo and Ma [2001]	Own Dataset 25 Images	HMM	…	187 Handwritten Words	92.86% precision,72.19% recall	English
Zheng *et al.* [2004]	Tobacco Industry Litigation Archives	Fisher Classifier	224 Documents	94 Documents	98.1% accuracy	English
Farooq *et al.* [2006]	Self-made Dataset	EM based NN	29 Documents	5 Documents	94.62% precision	Arabic
da Silva *et al.* [2009]	IAM, Forms	WEKA tool	…	1404 (P), 2029 (H) words	97.51%(P) 97.54%(H) for IAM, 97.17% (P), 99.47% (H)	English
Zemouri and Chibani [2011]	IAM	SVM	21 Documents	21 Documents	98% accuracy	English
Belaïd *et al.* [2013]	Industrial Documents	Multiclass SVM	75 Documents	300 Documents	90% accuracy	English
Awal *et al.* [2014]	ITESOFT	SVM	107 Documents	202 Documents	98.9% accuracy	English
Echi and Saidani [2014]	IAM, IFN-ENIT	Bayes AODEsr classifier	…	…	98.72% precision	Arabic, English
Ricquebourg *et al.* [2014]	Maurdor campaign Dataset	Adaboost (Bonzaiboost)	500,000 words	181,239 words	95% precision	English, French
Malakar *et al.* [2013]	Self-made Dataset (2000 words)	Decision Tree with feature ranking	250(H) + 250(P) words	750(H) + 750(P) words	96.80% accuracy	English

2.5 CLASSIFICATION AT CONNECTED COMPONENT LEVEL

Simple methods with lesser preprocessing overheads are generally preferable. Connected Component (CC) analysis is preferred for high-speed real time segmentation and recognition systems. A patch is defined to be a region in a document such that if a rectangular window (size determined dynamically or statically) is drawn with each foreground pixel within the patch as its center, then the window shall not contain any foreground pixel from another patch. Figure 2.3 demonstrates examples of such documents in real environment where mixed types of text persists and encourages segmentation at connected component level. Table 2.5 summarizes the connected component based handwritten and printed text classification methods in chronological order.

In one of the earliest work, [Franke and Oberlander, 1993] reported a method to check whether a CC is printed or handwritten. The height, width, gap and center-distance histogram features are extracted from within the minimum bounding box of a CC. Every histogram feature is passed individually to a polynomial classifier and its output is recorded. Finally, majority voting on the classifier outputs predicts the label for the CC.

Srihari *et al.* [1996] worked for automated forms processing for reading names and addresses on tax forms of the U.S. Internal Revenue Service. Fisher's classifier was used to discriminate hand-print/machine-print CC. For each CC, six distinctive features are computed, such as standard deviation of connected components width and height, average component density, aspect ratio, height and width. The classifier is trained for discrimination on 11013 documents and was tested on 800 postal address documents with a correct discrimination rate of 95%.

Santos *et al.* [2002] suggested to use content related features and shape related features to characterize handwritten text on bank check images. To extract the CC, a fixed size frame was used. All the features were fed into an MLP and label for each CC is predicted. In order to improve the misclassification rate, post processing based on neighborhood was applied

In a contemporary work by Kavallieratou *et al.* [2002], decision rules were applied to classify the CC's. For each CC a set of geometrical and structural features are extracted. The extracted feature values are sequentially tested against a set of four decision rules and accordingly decision labels are predicted. Their method achieves an accuracy of 96% on IAM forms.

Jang *et al.* [2004] used an MLP classifier to classify the machine-printed and handwritten addresses on images of Korean mail. The address image is mined to select valid CC by merging the smaller CC to form larger CC. Using geometric features extracted from each valid CC and training an MLP yielded an accuracy of nearly 99% for 3,147 testing images.

Likforman *et al.* [2006] addressed the problem of automatic extraction of names from UW English Document Database and facsimile images collected within the Majordome Project (2000-2003). Classification of printed and handwritten text is their pre-processing step. In their approach, morphological operations are used to extract CC's and termed as pseudo words. A set of statistical features for a pseudo word image is computed and used to train a neural network for classification. The work reported an overall accuracy of 77.2%.

As the electronic age is progressing, camera enabled devices have become popular and therefore instead of conventional scanners, cameras are commonly used to capture document images. A major problem in using cameras is the limited resolution and background noise that makes it difficult to process the electronic document for further OCR processing. To address this challenge, [Shetty *et al.*, 2007] used neighboring context to automatically label extracted patches as machine printed, handwritten or noise. A simple region growing algorithm was used to extract relevant patches from the document. Labels for these patches were inferred using a CRF model with Gibbs sampling. The model parameters are estimated by maximum pseudo-likelihood while the model is learnt using conjugate gradient descent. The feature set is built by a set of structural features and neighborhood features. The CRF model produced an accuracy of 95.75% on 27 documents of Tobacco dataset.

The work stated in [Kandan *et al.*, 2007] describes a two level classification algorithm to discriminate the handwritten elements from the printed text. At the first level the CC's are extracted and seven invariant central moments are computed. These feature are used to train an SVM and a NN classifier. At the second level, the labels predicted by these classifiers are post processed for better classification. Delaunay triangulation is used to reclassify the mis-classified elements. The method reports an accuracy of 87.85% when using NN, and 83.22% when using SVM. The proposed technique is independent of size, slant, orientation and translation in handwritten text.

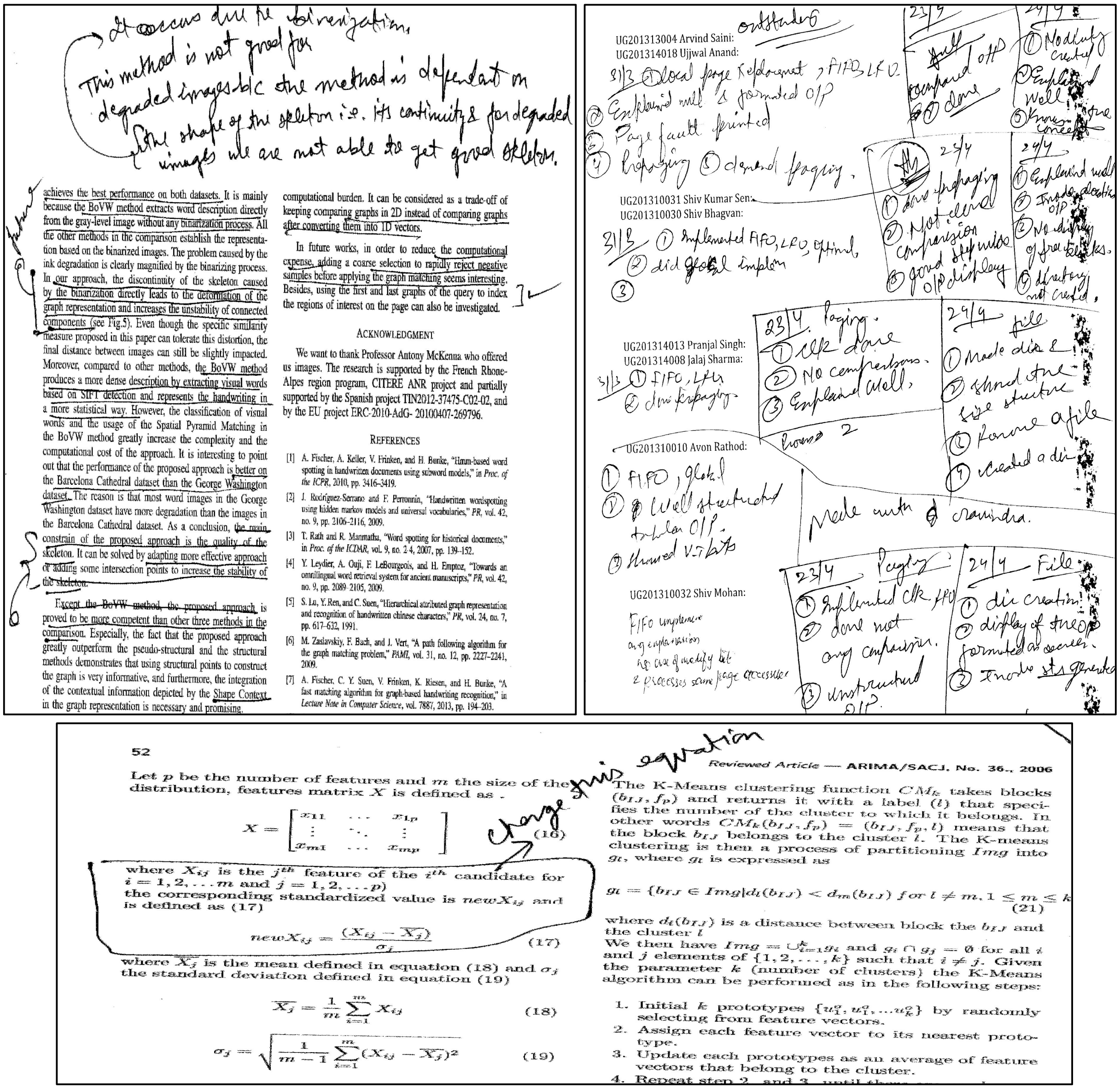
The task of annotation extraction gets difficult when the text content is sparse in the document. In this context, Chanda *et al.* [2010] proposed a method that achieved promising results of 96.90% accuracy, even when the document contains sparse data with arbitrary orientation. Prior to CC extraction, some preprocessing steps like region growing, and angle estimation using Principle Component Analysis are performed in order to resolve the arbitrary orientation issue. A chain-code histogram feature is used with an SVM classifier.

The work by Pinson and Barrett [2011] used template-based approach to separate the printed and handwritten text in administrative forms. Their work is inspired by Eigenfaces algorithm that is used for font and character recognition. They created 5,957 representative templates of fonts and styles of characters, numbers, punctuation and symbols. From these templates a hyper-dimensional character space is constructed from SVD and eigen value computations. Each extracted connected component is projected on to this space and is classified as printed or handwritten on the basis of a thresholded distance value. The method recorded 98% machine print precision and 28.95% handwriting precision. However, misclassification arises due to different font size and overlay printed text.

Benjlaiel *et al.* [2014] separated multi-oriented handwritten textual annotations based on internal and external shape analysis of a connected component. They used Gabor features, invariant moments and Fourier transform to capture shape and let the k-NN classifier to decide the category of the CC. Their method produces an accuracy of 98.48% for 301 documents.

With the increase in the application of codebook methods, [Barlas *et al.*, 2014] applied codebooks to classify handwritten and typed CC. For codebook generation, the external contour of the CC is extracted and small segments of fixed length are drawn. These segments are termed as fragments. These extracted fragments are represented by a chaincode histogram (CCH) and using them a codebook is created from a 2D Self-Organizing Map (SOM). The descriptors formed during vector quantization are fed into an MLP classifier which will predict the category of the CC. The codebooks are generated with Arabic and Latin CCs. For the codebook construction IFNENIT Arabic and RIMES Latin databases are used and for training the MLP classifier the MAURDOR dataset in Arabic and Latin is accessed. Their method achieved a precision of 79.0% for printed and 80.7% for handwritten. In the end, the labeled RLSA is applied to merge the CCs and form blocks of handwritten and printed data.

Annotations are not limited to text but there are many ways in which a document can be annotated. Pandey and Harit [2015] addressed the problem of multi-oriented handwritten text extraction in uncontrolled scenario. They geometrically segmented the complex cases of handwritten annotations, including marks, cuts and special symbols along, with the regular text. After morphologically CC extraction spectral partitioning is adopted as the segmentation scheme to separate the printed text and annotations. A new feature, called Envelope Straightness, was developed and included in the feature set along with other five statistical features. This led to an improvement of accuracy over the state-of-the-art features. In the framework of spectral partitioning, the addition of a new feature helped to achieve a recall of 98.39% for printed text and a precision of 85.40% for handwritten annotations on a dataset of 40 images (see Figure 2.4). IAM dataset has achieved a recall of 81.89% for printed text and a precision of 69.67% for handwritten annotations.



It occurs due to binarization.
This method is not good for degraded images b/c the method is dependant on the shape of the skeleton i.e. its continuity & for degraded images we are not able to get good skeleton.

future

achieves the best performance on both datasets. It is mainly because the BoVW method extracts word description directly from the gray-level image without any binarization process. All the other methods in the comparison establish the representation based on the binarized images. The problem caused by the ink degradation is clearly magnified by the binarizing process. In our approach, the discontinuity of the skeleton caused by the binarization directly leads to the deformation of the graph representation and increases the unstability of connected components (see Fig.5). Even though the specific similarity measure proposed in this paper can tolerate this distortion, the final distance between images can still be slightly impacted. Moreover, compared to other methods, the BoVW method produces a more dense description by extracting visual words based on SIFT detection and represents the handwriting in a more statistical way. However, the classification of visual words and the usage of the Spatial Pyramid Matching in the BoVW method greatly increase the complexity and the computational cost of the approach. It is interesting to point out that the performance of the proposed approach is better on the Barcelona Cathedral dataset than the George Washington dataset. The reason is that most word images in the George Washington dataset have more degradation than the images in the Barcelona Cathedral dataset. As a conclusion, the main constrain of the proposed approach is the quality of the skeleton. It can be solved by adapting more effective approach or adding some intersection points to increase the stability of the skeleton.

~~Except the BoVW method, the proposed approach~~ is proved to be more competent than other three methods in the comparison. Especially, the fact that the proposed approach greatly outperform the pseudo-structural and the structural methods demonstrates that using structural points to construct the graph is very informative, and furthermore, the integration of the contextual information depicted by the Shape Context in the graph representation is necessary and promising.

computational burden. It can be considered as a trade-off of keeping comparing graphs in 2D instead of comparing graphs after converting them into 1D vectors.

In future works, in order to reduce the computational expense, adding a coarse selection to rapidly reject negative samples before applying the graph matching seems interesting. Besides, using the first and last graphs of the query to index the regions of interest on the page can also be investigated.

ACKNOWLEDGMENT

We want to thank Professor Antony McKenna who offered us images. The research is supported by the French Rhone-Alpes region program, CITERE ANR project and partially supported by the Spanish project TIN2012-37475-C02-02, and by the EU project ERC-2010-AdG- 20100407-269796.

REFERENCES

[1] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Hmm-based word spotting in handwritten documents using subword models," in *Proc. of the ICPR*, 2010, pp. 3416–3419.

[2] J. Rodríguez-Serrano and F. Perronnin, "Handwritten wordspotting using hidden markov models and universal vocabularies," *PR*, vol. 42, no. 9, pp. 2106–2116, 2009.

[3] T. Rath and R. Manmatha, "Word spotting for historical documents," in *Proc. of the ICDAR*, vol. 9, no. 2-4, 2007, pp. 139–152.

[4] Y. Leydier, A. Ouji, F. LeBourgeois, and H. Emptoz, "Towards an omnilingual word retrieval system for ancient manuscripts," *PR*, vol. 42, no. 9, pp. 2089–2105, 2009.

[5] S. Lu, Y. Ren, and C. Suen, "Hierarchical attributed graph representation and recognition of handwritten chinese characters," *PR*, vol. 24, no. 7, pp. 617–632, 1991.

[6] M. Zaslavskiy, F. Bach, and J. Vert, "A path following algorithm for the graph matching problem," *PAMI*, vol. 31, no. 12, pp. 2227–2241, 2009.

[7] A. Fischer, C. Y. Suen, V. Frinken, K. Riesen, and H. Bunke, "A fast matching algorithm for graph-based handwriting recognition," in *Lecture Note in Computer Science*, vol. 7887, 2013, pp. 194–203.

UG201313004 Arvind Saini: outstanding
UG201314018 Ujjwal Anand:
31/3 ① local page Replacement, FIFO, LFU.
② Explained well & formated O/P
③ Page fault printed
④ Prepaging ⑤ demand paging.

23/4 full compared O/P ① done

24/4 ① Modularity created ② Explained well ③ known the concept

UG201310031 Shiv Kumar Sen:
UG201310030 Shiv Bhagvan:
31/3 ① Implemented FIFO, LRU, Optimal,
② did global implem
③

23/4 ① done prepaging ② Not done comparision good step wise ③ O/P display

24/4 ① Explained well ② Inode details O/P ③ No display of free blocks. ④ directory not created.

UG201314013 Pranjal Singh:
UG201314008 Jalaj Sharma:
31/3 ① FIFO, LFU
② done prepaging

23/4 Paging. ① clk done ② No comparisons. ③ Explained well.

24/4 file ① Made dir & ② shared the size structure ③ Remove a file ④ created a dir

Process 2

UG201310010 Avon Rathod:
① FIFO, global
② Well structured tabular O/P.
③ showed V-T bits

Made with Ravindra.

UG201310032 Shiv Mohan:
FIFO implemen
any explanation
no case of modify bit
2 processes same page accessed

23/4 Paging ① Implemented clk, FIFO ② done not any comparison. ③ unstructured O/P.

24/4 File ① dir creation ② display of the op formated as screen. ③ Inode sts generated

52

Let p be the number of features and m the size of the distribution, features matrix X is defined as .

$$X = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mp} \end{bmatrix} \quad (16)$$

change this equation

where X_{ij} is the j^{th} feature of the i^{th} candidate for $i = 1, 2, \dots m$ and $j = 1, 2, \dots p$)
the corresponding standardized value is $newX_{ij}$ and is defined as (17)

$$newX_{ij} = \frac{(X_{ij} - \overline{X_j})}{\sigma_j} \quad (17)$$

where $\overline{X_j}$ is the mean defined in equation (18) and σ_j the standard deviation defined in equation (19)

$$\overline{X_j} = \frac{1}{m}\sum_{i=1}^{m} X_{ij} \quad (18)$$

$$\sigma_j = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(X_{ij} - \overline{X_j})^2} \quad (19)$$

Reviewed Article — ARIMA/SACJ, No. 36., 2006

The K-Means clustering function CM_k takes blocks (b_{IJ}, f_p) and returns it with a label (l) that specifies the number of the cluster to which it belongs. In other words $CM_k(b_{IJ}, f_p) = (b_{IJ}, f_p, l)$ means that the block b_{IJ} belongs to the cluster l. The K-means clustering is then a process of partitioning Img into g_l, where g_l is expressed as

$$g_l = \{b_{IJ} \in Img | d_l(b_{IJ}) < d_m(b_{IJ}) \; for \; l \neq m, 1 \leq m \leq k \quad (21)$$

where $d_l(b_{IJ})$ is a distance between block the b_{IJ} and the cluster l
We then have $Img = \cup_{i=1}^{k} g_i$ and $g_i \cap g_j = \emptyset$ for all i and j elements of $\{1, 2, \dots, k\}$ such that $i \neq j$. Given the parameter k (number of clusters) the K-Means algorithm can be performed as in the following steps:

1. Initial k prototypes $\{u_1^o, u_1^o, \dots u_k^o\}$ by randomly selecting from feature vectors.
2. Assign each feature vector to its nearest prototype.
3. Update each prototypes as an average of feature vectors that belong to the cluster.
4. Repeat step 2 and 3 until there

Figure 2.4 : Multi-oriented Handwritten Dataset used by the work [Pandey and Harit, 2015].

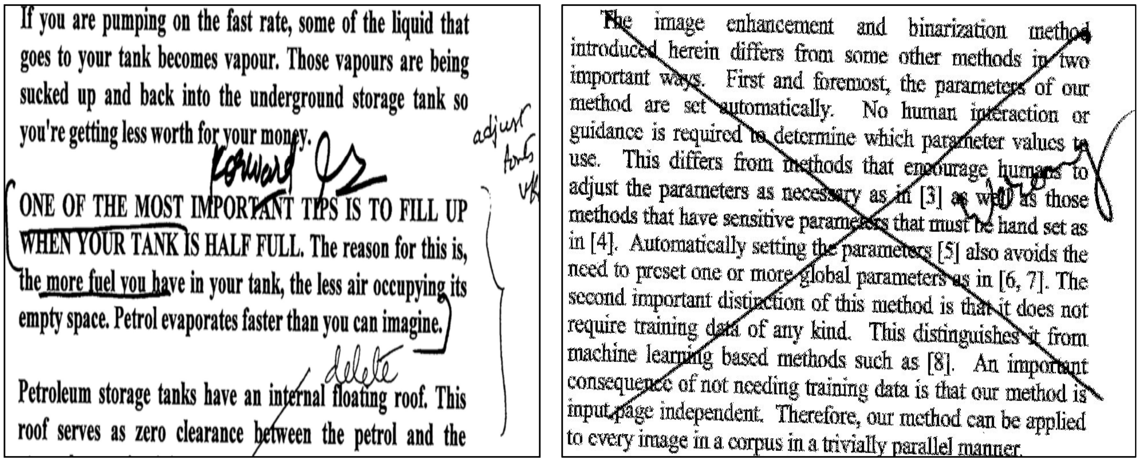
The work presented by [Awal and Belaïd, 2017] addressed the problem of handwritten and printed text separation in Arabic document images. After document cleaning and orientation correction, the CC called pseudo words were extracted. A total of 137 features were extracted from the pseudo-word. These include statistical and geometric features. A local classification step, using a Gaussian kernel SVM, associated each pseudo-word into either handwritten or printed class. During post processing the labels propagate in the pseudo-word's neighborhood in order to recover from classification errors. The proposed methodology achieved a separation rate of around 90%.

Table 2.5 : Overview of dataset, performance and methods applied for classification of handwritten connected components from printed connected components.

System	Dataset	Segmentation/ Classification Scheme	Training Data	Testing Data	Performance	Language
Franke and Oberlander [1993]	Mails	Polynomial Classifier	4000 CC	4000 CC	5% error	English
Srihari *et al.* [1996]	Tax forms	Fisher's classifier	11013 documents	800 Documents	95%	English
Santos *et al.* [2002]	70 Brazilian Banks cheques Images	MLP	6772 CC	5035 CC	88% accuracy	English
Kavallieratou *et al.* [2002]	IAM Dataset	Decision rules	…	…	96%	English
Jang *et al.* [2004]	Korean Mail Piece	MLP	9,028 CC	3,147 CC	98.9%	Korean
Likforman *et al.* [2006]	UW English Document Database facsimile	Neural Network	960 words Documents	…	77.2% accuracy	French, English
Kandan *et al.* [2007]	Official Documents	SVM, NN classifier	350 documents	1,678 handwritten CC	87.85% with NN, 83.22% with SVM	English
Shetty *et al.* [2007]	Tobacco industrial	CRF	3700 patches	3800 patches	95.75%	English
Chanda *et al.* [2010]	Self-made dataset	SVM	1500 Documents	…	96.90%	Roman script
Pinson and Barrett [2011]	NIST SD19	Eignface algorithm	…	360 binary images	71.05% for handwritten and 98.21% for printed	English
Benjlaiel *et al.* [2014]	Self-made Dataset	k-NN	…	301 documents	98.48%	English
Barlas *et al.* [2014]	MAURDOR Dataset	Codebook with MLP	(25000 × 3) samples	…	79.0% (P), 80.7% (H) precision, 83.6% (P), 75.7% (H) recall	French, English, Arabic
Pandey and Harit [2015]	Self-made, IAM	Spectral Partitioning	…	40 documents	98.39% (P), 85.40% (H) precision	English
Awal and Belaïd [2017]	Arabic real dataset	SVM	…	…	90%	Arabic

2.6 CLASSIFICATION AT PIXEL LEVEL

Pixels are the basic units of images. While annotating a document it may happen that the hand-marked annotations get overlapped on the printed text. For such cases, separation at pixel level is desirable. Figure 2.5 demonstrates examples of such documents where annotations overlap with the text and call for a segmentation of handwritten and printed content at pixel level.



If you are pumping on the fast rate, some of the liquid that goes to your tank becomes vapour. Those vapours are being sucked up and back into the underground storage tank so you're getting less worth for your money.

forward

adjust fonts

ONE OF THE MOST IMPORTANT TIPS IS TO FILL UP WHEN YOUR TANK IS HALF FULL. The reason for this is, the more fuel you have in your tank, the less air occupying its empty space. Petrol evaporates faster than you can imagine.

delete

Petroleum storage tanks have an internal floating roof. This roof serves as zero clearance between the petrol and the

The image enhancement and binarization method introduced herein differs from some other methods in two important ways. First and foremost, the parameters of our method are set automatically. No human interaction or guidance is required to determine which parameter values to use. This differs from methods that encourage humans to adjust the parameters as necessary as in [3] as well as those methods that have sensitive parameters that must be hand set as in [4]. Automatically setting the parameters [5] also avoids the need to preset one or more global parameters as in [6, 7]. The second important distinction of this method is that it does not require training data of any kind. This distinguishes it from machine learning based methods such as [8]. An important consequence of not needing training data is that our method is input page independent. Therefore, our method can be applied to every image in a corpus in a trivially parallel manner.

Figure 2.5 : Examples of such documents [Peng *et al.*, 2013; Pandey and Harit, 2015] in real environment where the text is of overlapping nature and encourages segmentation of handwritten and printed content at Pixel Level.

Table 2.6 presents a tabulated summary of the work on classification of printed and handwritten pixels in chronological order.

Nakai *et al.* [2007] developed a method based on alignment matching for separating out the marginal annotations. The document is scanned in its original form, and then again after it has been annotation. The annotations are extracted by point matching the annotated document with the original document. On a dataset of 327 document images, the marginal annotations were extracted with a precision of 78%.

Nakai *et al.* [2008] also used alignment matching, but were able to improve annotation extraction by removing image degradation while scanning. Annotations were extracted by subtracting the original document image from the annotated document image. A major problem of annotation extraction is that images suffer from slight displacements even after doing local alignment. Rough handling can contaminate an annotated document by adding noise to it. Moreover, the noise reduction step can have the side effect of removing noise as well as annotations. Alignment matching is difficult in such situations and instead of using common alignment matching parameters, its better to use different parameters for different areas. In the method proposed by Nakai *et al.* [2008], the input image is divided into several patches and the alignment parameters are computed. For matching purpose, the parameter value for a patch is computed as the median of the four neighborhood values. This reduces errors due to distortion or change in alignment and produces a recall of 80.94% and precision of 85.59%.

Peng *et al.* [2013] suggested to use coherent pixel aggregation scheme to separate handwritten text from the overlayed printed text. The contrast of a handwritten pixel in context with the neighborhood pixels is captured by shape context feature [Belongie *et al.*, 2002]. The overlapped text image is modeled by a graph where pixels are connected to their nearest neighbors. The neighboring pixels are clustered based on the variance in their shape descriptors. The process continues for the growing pixel segments till a pre-defined threshold. To further improve the results, MRF is applied as a post processing scheme and this yields an accuracy of 86.82%.

For overlay text separation [Seuret *et al.*, 2014] suggested a window-based method to classify the pixels. Feature extraction was done for each window and a large set of statistical and geometric features were extracted at every pixel from its square neighborhood. The pixels weree classified as printed or handwritten by a standard MLP. The final result was rectified by applying post processing techniques based on heuristics. The dataset comprised 102 high quality colored documents annotated by 15 different authors in Chinese, Cyrillic, English, Arabic and Latin. The whole system was trained and validated with 78 documents and the remaining were used for testing. It reached an accuracy of 97.70% for validation data and 96.10% with test data. The best results were obtained for Chinese and Cyrillic, while for Latin the results were comparatively inferior.

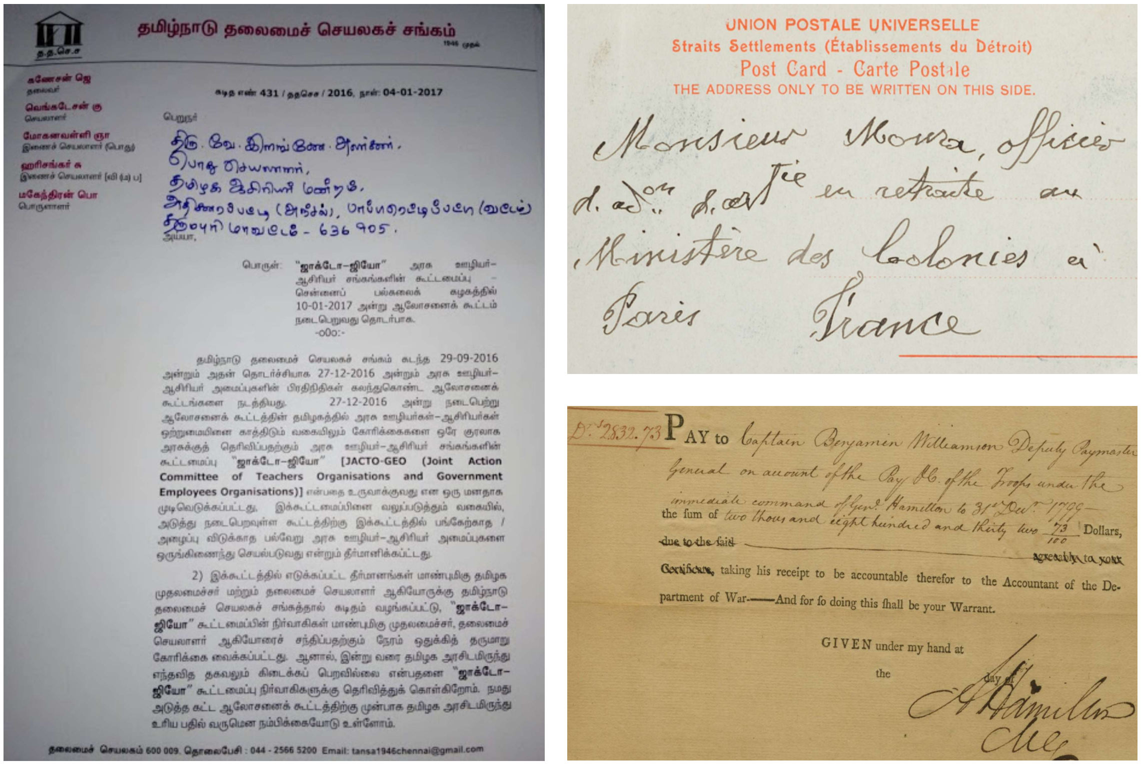
Table 2.6 : Overview of Dataset, Performance and Methods applied for Classifying Handwritten Pixels from Printed Pixels.

System	**Dataset**	**Segmentation/ Classification Scheme**	**Training Data**	**Testing Data**	**Performance**	**Language**
Nakai *et al.* [2007]	327 Documents from Magazines and Newspapers	Point Matching	…	327 Documents	78%	Chinese, Japanese
Nakai *et al.* [2008]	self-created Dataset	Alignment Matching	…	100 pages	recall 80.94%, precision 85.59%	Japanese
[Peng *et al.*, 2013]	Self-created over-lapped dataset	Aggregation coarsening	110 patches	110 patches	86.82%	English
Seuret *et al.* [2014]	PRImA	MLP	1000000 patches	500000 patches	96.10%	Chinese, Cyrillic, English, Arabic, Latin

2.7 CLASSIFICATION AT LINE LEVEL

The need for OCR in business, academic, and personal usage is increasing rapidly. OCRs are now expected to be fast and efficient. Digitizing the handwritten content is the need of the era and this requires segmenting the handwritten content from the printed text line by line and then recognizing it. Figure 2.6 demonstrates examples of such documents where there are long text-lines of single category (i.e. containing all handwritten or all printed text) and this encourages segmentation of handwritten and printed content at text-line level. Table 2.7 presents abridged description for the classification of printed and handwritten text-lines in chronological order.

Printed characters posses regularity in arrangement. On the other hand, the handwritten



தமிழ்நாடு தலைமைச் செயலகச் சங்கம்
1946 முதல்
த.த.செ.ச

கணேசன் ஜெ
தலைவர்
வெங்கடேசன் கு
செயலாளர்
மோகனவள்ளி ஞா
இணைச் செயலாளர் (பொது)
ஹரிசங்கர் சு
இணைச் செயலாளர் [வி (ம) ப]
மகேந்திரன் பொ
பொருளாளர்

கடித எண் 431 / ததசெச / 2016, நாள்: 04-01-2017

பெறுநர்

திரு. வே. இளங்கோ. அவர்கள்.
பொது செயலாளர்,
தமிழக ஆசிரியர் மன்றம்.
அதிகாரப்பட்டி (அஞ்சல்), பாப்பிரெட்டிப்பட்டி (வட்டம்)
தர்மபுரி மாவட்டம் - 636 905.

அய்யா,

பொருள்: "ஜாக்டோ–ஜியோ" அரசு ஊழியர்–ஆசிரியர் சங்கங்களின் கூட்டமைப்பு – சென்னைப் பல்கலைக் கழகத்தில் 10-01-2017 அன்று ஆலோசனைக் கூட்டம் நடைபெறுவது தொடர்பாக.

-o0o:-

தமிழ்நாடு தலைமைச் செயலகச் சங்கம் கடந்த 29-09-2016 அன்றும் அதன் தொடர்ச்சியாக 27-12-2016 அன்றும் அரசு ஊழியர்–ஆசிரியர் அமைப்புகளின் பிரதிநிதிகள் கலந்துகொண்ட ஆலோசனைக் கூட்டங்களை நடத்தியது. 27-12-2016 அன்று நடைபெற்ற ஆலோசனைக் கூட்டத்தின் தமிழகத்தில் அரசு ஊழியர்கள்–ஆசிரியர்கள் ஒற்றுமையினை காத்திடும் வகையிலும் கோரிக்கைகளை ஒரே குரலாக அரசுக்குத் தெரிவிப்பதற்கும் அரசு ஊழியர்–ஆசிரியர் சங்கங்களின் கூட்டமைப்பு **"ஜாக்டோ–ஜியோ" [JACTO-GEO (Joint Action Committee of Teachers Organisations and Government Employees Organisations)]** என்பதை உருவாக்குவது என ஒரு மனதாக முடிவெடுக்கப்பட்டது. இக்கூட்டமைப்பினை வலுப்படுத்தும் வகையில், அடுத்து நடைபெறவுள்ள கூட்டத்திற்கு இக்கூட்டத்தில் பங்கேற்காத / அழைப்பு விடுக்காத பல்வேறு அரசு ஊழியர்–ஆசிரியர் அமைப்புகளை ஒருங்கிணைந்து செயல்படுவது என்றும் தீர்மானிக்கப்பட்டது.

2) இக்கூட்டத்தில் எடுக்கப்பட்ட தீர்மானங்கள் மாண்புமிகு தமிழக முதலமைச்சர் மற்றும் தலைமைச் செயலாளர் ஆகியோருக்கு தமிழ்நாடு தலைமைச் செயலகச் சங்கத்தால் கடிதம் வழங்கப்பட்டு, **"ஜாக்டோ–ஜியோ"** கூட்டமைப்பின் நிர்வாகிகள் மாண்புமிகு முதலமைச்சர், தலைமைச் செயலாளர் ஆகியோரைச் சந்திப்பதற்கும் நேரம் ஒதுக்கித் தருமாறு கோரிக்கை வைக்கப்பட்டது. ஆனால், இன்று வரை தமிழக அரசிடமிருந்து எந்தவித தகவலும் கிடைக்கப் பெறவில்லை என்பதனை **"ஜாக்டோ–ஜியோ"** கூட்டமைப்பு நிர்வாகிகளுக்கு தெரிவித்துக் கொள்கிறோம். நமது அடுத்த கட்ட ஆலோசனைக் கூட்டத்திற்கு முன்பாக தமிழக அரசிடமிருந்து உரிய பதில் வருமென நம்பிக்கையோடு உள்ளோம்.

தலைமைச் செயலகம் 600 009. தொலைபேசி : 044 - 2566 5200 Email: tansa1946chennai@gmail.com

UNION POSTALE UNIVERSELLE
Straits Settlements (Établissements du Détroit)
Post Card - Carte Postale
THE ADDRESS ONLY TO BE WRITTEN ON THIS SIDE.

Monsieur Moura, officier
d. adm. de 1.re en retraite au
Ministère des Colonies à
Paris France

Drs 2832.73 PAY to Captain Benjamin Williamson Deputy Paymaster General on account of the Pay &c. of the Troops under the immediate command of Genl. Hamilton to 31st Decr. 1799 the ſum of two thousand eight hundred and thirty two $\frac{73}{100}$ Dollars, ~~due to the ſaid~~ ~~agreeably to your Certificate,~~ taking his receipt to be accountable therefor to the Accountant of the Department of War.——And for ſo doing this ſhall be your Warrant.

GIVEN under my hand at
the day of

A Hamilton

Figure 2.6 : Examples of such documents [Lin Yangchen, 2015; Nathan Raab, 2017] where long independent text-lines of single category persist and encourages segmentation of handwritten and printed content at text-line Level.

characters, more often, do not depict regularity in arrangement. Based on this fact, Fan *et al.* [1998] proposed a method to classify and separate the handwritten text lines from printed text. Each text line is divided into their corresponding characters or sets of characters using XY cuts. These blocks are termed as character blocks. These blocks are then classified as handwritten and printed character blocks based on the degree of straightness. The center bottom point for each character block in the text line is located. The distribution of these points aids to compute the degree of straightness in a text line. It is observed that the distribution of such points in character blocks is regular for machine-printed texts, and irregular for handwritten texts. This feature is more stable for a variety of languages and termed as character block layout variance. The experiment was set on 25 handwritten and printed English and Chinese documents in 7 different fonts. The accuracy achieved was about 86%.

Pal and Chaudhuri [1999] addressed the need of multi-script OCR for Indian languages. A tree-based classification approach is used to segment handwritten and printed text lines of Bangla and Devnagari. In the preprocessing step, RLSA along with histogram based thresholding is applied to segment text-lines. A set of upper and lower profile features are extracted for each text-line and given to a tree classifier to label them as printed and handwritten. It reports an accuracy of 98.3% on 100 images from question papers, money order form, application form, letter etc., in Bangla and Devanagari. The given approach is size and font independent. However, for short text-lines with one or two words the scheme produced incorrect results.

In extension to the previous work, [Pal and Chaudhuri, 2001] proposed a feature that separates machine-printed and hand-written text lines. The new feature was inspired from [Fan *et al.*, 1998]. This feature could find the straightness of a given text-line. It is termed as character lowermost point standard deviation (CLPSD). The classification method had two steps. The first step used a set of statistical features to classify a text line as handwritten or ambiguous. The next step further classified an ambiguous line as printed or handwritten using CLPSD features. This method was tested on 600 different documents of Bangla and Devanagari and achieved an accuracy 98.6%. Due to script similarity, their method can be applied to other Indian languages also, such as Marathi, Assamese, and Punjabi.

Kavallieratou and Stamatatos [2004] also exploited the structural properties of printed and handwritten text-lines. Images are skew corrected using the Wigner-Ville distribution (WVD) [Boashash and O'Shea, 1994] and text lines are located by run-length smearing algorithm. For every text-line the ascender, descender and core zones were marked and based on them three sets of features were extracted. Discriminant analysis with Mahalonobis distance was used to predict for unseen cases. The experiment was performed on two datasets: IAM-DB (English text) and GRUHD (Greek text) for 50 documents and a classification accuracy of 98.2% was achieved. However, the method misclassified short handwritten text lines with usually one or two words.

Table 2.7 : Overview of Dataset, Performance and Methods applied for Handwritten Text-line Classification from Printed Text-line.

System	Dataset	Segmentation/ Classification Scheme	Training Data	Testing Data	Performance	Language
Fan *et al.* [1998]	25 Documents from Magazines and Newspapers	Heuristic based method	…	25 Documents	86%	Chinese, English, Japanese
Pal and Chaudhuri [1999]	Self-made Dataset	Tree based Classifier	…	100 Documents	98.3%	Bangla, Devanagari
Pal and Chaudhuri [2001]	600 Documents	Heuristic based method	…	600 Documents	98.6%	Bangla and Devanagari
Kavallieratou and Stamatatos [2004]	IAM-DB, GRUHD	Discriminant Analysis	50 Documents	…	98.2%	English, Greek

2.8 IMPROVING THE CLASSIFICATION RATE: POST PROCESSING

The classification performance gets drastically affected by the presence of artifacts when categorizing the text as printed or handwritten. A major trend of research in this area is to first segment the document into individual basic units, classify them into intended categories, and then do post processing, i.e., reconcile this classification by a neighborhood technique to recover from classification errors. Single errors in the classification of text as handwritten or printed can be corrected by inspecting and reconciling with the neighbourhood word. This means that if a word classified as handwritten is nested in a machine printed text line and its neighbors are machine printed, then this word will be relabeled as machine printed. On the other hand, if a word classified as machine printed has no line forming information and is surrounded by handwritten words, it will be re-labeled as handwritten text. Table 2.8 lists the post processing methods along with their contributions in performance improvement for handwritten vs printed text classification. The work reported by [Guo and Ma, 2001; Santos *et al.*, 2002; Shetty *et al.*, 2007] applied the above neighborhood post-processing to rectify the results.

Due to unconstrained annotation style, it may not be necessary that we always have straightness of lines in our text. For such scenarios, there is need of post processing methods that dynamically find their neighbors. Examples of such schemes are Contextual re-labeling [Zheng *et al.*, 2004; Belaïd *et al.*, 2013; Seuret *et al.*, 2014; Awal *et al.*, 2014; Awal and Belaïd, 2017] and Delaunay triangulation [Kandan *et al.*, 2007].

2.8.1 Delaunay Triangulation

Delaunay triangulation [Davoine *et al.*, 1996] of a set of non-degenerate vertices is defined as the unique triangulation with empty circles, i.e, no vertex lies inside the circumscribing circle of any Delaunay triangles. Delaunay triangulation carried out on printed or handwritten text has the following properties:

1. The lengths of the sides of most triangles in a printed text region are similar as compared to the lengths of the handwritten text.

2. Triangles in the printed text have their longest and similar sides link the point pairs separating two adjacent text lines.

3. The heights of the triangles in the printed text region are uniform.

2.8.2 Contextual Re-labeling

Contextual relabeling is a two-step process:

1. Identifying the neighbors of the pseudo-word
2. Re-labeling

Awal *et al.* [2014] stated that contextual re-labeling is operated by three different grouping techniques:

1. *k*-NN : It is based on searching of the *k* nearest neighbors. If more than 50% of neighbors share the same label, then the majority label is assigned to the central component. This majority is decided by a pre-defined threshold supplied manually.

2. k-NN with constraints: Some constraints are added to avoid small components to interfere with the label updating. Hence, before flipping the pseudo-word label, a test is performed to check whether the accumulated number of pixels in the neighbors is significant compared to the number of pixels of the main component.

3. Confidence Propagation: At times there is an isolated handwritten component which needs no alteration to its label. To avoid such random updates, the classifier confidence of the nearest horizontal neighbor is used. If the neighbor is stronger than that of the pseudo-word, the neighborhood class is assigned. A Gaussian function weighs the neighbor confidence by its distance to the pseudo-word.

Awal *et al.* [2014] proposed three post-processing methods which were used by [Farooq *et al.*, 2006] to improve the classification results for Arabic. They proposed three methods:

1. CRF relabeling: CRF is directly applied on pseudo words to estimate the classification confidences of each pseudo word. The neighborhood of a pseudo word comprises its left/right horizontal neighbors. The output probabilities of the local classifier and structural features like height ratio, position ratio and density ratio from neighborhood are selected as features to train the CRF.
2. Grouping by pseudo-lines (CRF Probabilistic Method) Grouping of the pseudo lines is done by modeling the logical relationship between its pseudo-words, in addition to the spatial relationships along the horizontal direction. For the probabilistic model, CRF is trained with features such as: Height ratio, Density ratio, CC count ratio and Inter-CC distance variance ratio.
3. Grouping by pseudo-lines (Determinstic Method) In the deterministic model the median height for the entire pseudo line is calculated and compared with each pseudo word of the same pseudo text line.

Table 2.8 : Overview of Post Processing Methods applied for Classification of Handwritten Text Printed Text.

System	Post Processing	Performance Improvement Rate	Language
Guo and Ma [2001]	Neighborhood Analysis using Threshold	…	English
Santos *et al.* [2002]	Neighborhood Analysis using Threshold	…	English
Zheng *et al.* [2004]	MRF contextual relabeling	2.1%	English
Kandan *et al.* [2007]	Delaunay triangulation	…	English
Shetty *et al.* [2007]	Neighborhood Analysis using Threshold	…	English
Song *et al.* [2011a]	MRF	…	English, Chinese, Japanese
Belaïd *et al.* [2013]	kd trees contextual relabeling	…	English
Awal *et al.* [2014]	CRF contextual relabeling	1.3%	English
Seuret *et al.* [2014]	CRF contextual relabeling	1.2%	English
Awal and Belaïd [2017]	Pseudo-line based extension	5.1%	Arabic

2.9 CONCLUSION

In this chapter, the *state-of-the-art* methods for printed and handwritten text classification are reviewed. The distinguishing properties of hand printed and machine printed text are illustrated in Table 2.1. These characteristics have influenced the design of different features to discriminate between the two classes. Although, few methods are there that have considered classifying overlay text, it is found that most of the literature concerns with the problem of classifying well separated text as printed and handwritten. Most of the existing work has confined to considering handwritten annotations as textual and therefore restricted the analysis to characters, words, lines, or connected components. But, in the general case, handwritten annotations can include anything written or drawn by hand. They can be cuts, crosses, arrows, underlines, pictures, flowcharts, inline text, and special symbols. Hence, there is a need to consider other annotations also, apart from only handwritten text. There is a need to develop new methods and features that can categorize the annotations and can extract annotated regions of, say, only a specific *type* of annotation. Following this, in the subsequent Chapters 4, 5 and 6 we present proposals of new features and methods to distinguish printed text from complex annotations. In the next chapter we present a survey of the past work on off-line writer identification.

…