Classification of Document Image Components

B.V.Dhandra
Dept. of Computer Science, Gulbarga University, Gulbarga

Shridevi Soma
Dept. of CSE, PDA College of Engineering, Gulbarga

RashmiTallalli
Dept. of CSE, PDA College of Engineering, Gulbarga

GururajMukarambi
Dept. of Computer Science, Gulbarga University, Gulbarga

Abstract

A method is devised to classify the underlying document image components such as handwritten text, printed text, seal, graphs, tables etc. so as to address the problem of indexing and retrieving of the document images. The scanned document image is processed and segmented using the rule based method and connected component labelling to isolate the distinct image entities. The shape features such as area, perimeter, form factor, major axis, minor axis, roundness, compactness, density, white pixels of each line and vertical projection variance are extracted. The k – nearest neighbour classifier is used to characterise the distinct image entities of the document.

The experimental result exhibits the efficiency of the proposed system as 91.057% on an average.

Index terms – Connected Component Labelling, Document Images, Word Segmentation, Printed and Handwritten text Recognition, Seal Identification

1. Introduction

Normally a document is a paper that contains the printed and handwritten text component along with tables, graphs, stamps, seals, logos, circuit diagrams, pictures etc. A document may be in simple or complex form. Now-a-days a wide variety of information which is stored on paper is converted into digital form through a scanner or a fax machine for efficient storage and intelligent processing. Some of the common documents are application forms for college admission, business letters, bank cheques, challans, circulars, doctor's prescriptions, documents in postal department, engineering drawings and maps, exam hallticket, incomtax letters, invoices, petitions, requests, purchase bills, symbolic data, technical manuals etc. All these document images are processed using digital image processing techniques for potential information extraction, retrieval, modification, transmission and reuse.

The objective of the paper is to classify the components of the English document image that consists of textual and graphical components. The system is aimed at identifying the printed text, handwritten text and an image object like seal, logo, symbol etc. from the official document. In the proposed method circular shaped seal and the

English uppercase and lowercase characters are considered for identification and classification. The identification of handwritten text is a crucial task as compared to the machine printed text that is uniform in nature. The reference processing of complex documents is a challenging task for document application on live retrieval and recognition. The algorithms that work well on simple documents may not work well on complex documents (which are mixture of noise, handwriting, machine printed text with different fonts and font sizes, seals, tables, stamps, rule lines) as the elements impose a lot of restrictions to the algorithms. The primary task of processing such document images is to isolate the different components of the document. The documents after segregating the components, documents are known as indexed documents. Indexing of documents can be performed based on textual and/or graphical entities. Thepresent trend of organizations is to implement the digital mailrooms to enhance the efficiency of paper intensive workflows and to lessen the load of processing information of incoming mails, faxes, invoices, reports, etc. By this digital mailroom, organizations are forced for automatic distribution of incoming mails to their respective departments based on content of the digital documents.

The paper is organized as follows: Section 2 focuses on literature survey of the related work. Section 3 presents the detailed description of the proposed methodology. Section 4 describes the experimental results. Conclusion and future work are the subject matter of Section 5.

2. Literature Survey

Upasana. Patil et. al. (2012) proposed a method to discriminate the handwritten and printed text components based on shape features and word level separation. They have converted document image into binary image using Otsu's threshold from grey-level histograms and morphological operations are applied to remove the noise. Connected component processing is performed and they have observed that there are differences in the shape of the connected components of a handwritten and printed text words and these discriminating features are extracted using simple shape features such as circularity, major axis, minor axis etc. KNN classifier is used to classify the printed and handwritten text. They have obtained an accuracy of 98.57%. Konstantinos Z et. al. (2012) have developed a method to identify and separate the handwritten text from machine printed text using the bag of visual words paradigm (BoVW) and scale – invariant feature transform (SIFT). A visual word is represented as a group of features which correspond to the local image information which is identified by the image key points. The blocks of interest are detected initially in the image. A descriptor is calculated based on BoVW for every block of interest. The blocks are separated as handwritten, machine printed and noise by using a Support Vector Machine classifier. This was evaluated on two datasets namely IAM and PRImA-NHM and they have reported 98.86 and 76.89 accuracy.

T Kasar et. al. (2011) described a method for extraction of colour text from natural scene images. A two – fold smoothing of color pixels is performed row wise and column wise sequentially. Each edge segment is replaced by the median of the color values of pixels in that particular segment to obtain smoothed image. They have used the color information and stroke width of the connected components to identify the foreground components (text). They have tested on word images from the ICDAR 2003 robust reading competition dataset and obtained the recognition accuracy of 91.9% and is significant. Shirdhonkar et. al. (2010) have devised a system for automatic identification of the signature in scanned document images which helps to retrieve the document images based on signature. They have used the region growing algorithm to segment the document image into a number of segments(patches). Eight state features such as height, aspect ratio, maximum run length etc., are extracted from all the segments. A label is assigned to each segment using neural network (NN) andSupport Vector Machine (SVM) to classify printed and handwritten text. These models assume signature as a type of handwritten document. Recall and precision are the metrics used to evaluate the classification rate. The recall and precision values for machine printed text using SVM is 92 and 96 respectively, whereas using NN it is 86 and 96 respectively. The recall and precision values for handwritten signature using SVM is 100 and 75 respectively, whereas by using NN it is 100 and 50 respectively. The authors experimentally found that rate of classification of SVM is superior over NN. U. Pal et. al. (2010) proposed a method for seal detection in the document using generalized houghtransform and character proximity graphs. Using connected components, the rotation invariant spatial feature descriptors are computed and the support vector machine (SVM) classifier is used to recognize multi-oriented and multi-scale text characters. A voting procedure is designed to determine the location of the seal in a document. The accumulator of GHT records the votes to find location of seal. Experimental results prove the robustness of the proposed approach. The rectangular, circular and elliptical seal recognition accuracy are 93.61%, 91.74% and 87.50% respectively. Arun C et. al. (2013) presented a

method of connected component labelling and extraction based on interphase removal from the chromosome images. The connected components are identified and labelled. The components which have greater number of pixels than the predefined amount of pixels are obtained from the image which generates another image that contains purely the chromosome. The difference of segmented image andoriginal image is obtained and this in turn facilitates accurate karyotyping procedures. Their technique was tested on a standard clinical database of human karyotype and the segmentation accuracy achieved is 98.246%. S S Bukhari et. al. (2010) proposed a method for document image segmentation using discriminative learning process over connected components. The document image is segmented into text and non-text regions which is an important step in document image analysis. Connected component based classification approach is used. A self-tuneable multi-layer perceptron(MLP) is trained to classify the document image into text and non-text using shape and context information features. Their proposed method had been evaluated on UW-III(University of Washington-III) dataset, page segmentation test images and circuit diagrams dataset. For each dataset, pixel-level ground truth has been generated using zone-level ground truth information. Their evaluation results proved the effectiveness of the proposed method, exact accuracy is not available. Sargur.N.Srihari et. al. (2008) reported a new distance metrics based word segmentation algorithm. The local and global features are extracted and a three-layer neural network is used for classifying the inter-word distance. An unconstrained handwritten database is used to evaluate the system and they have shown the overall accuracy as 90.82%.

Thus from the above literature, it is clear that the classification of image documents either suffers from the classification accuracy or small feature set or from time complexity. Hence, there is a need toaddress this problem with respect to one of the above factors or in combination.

3. Document Image Classification

The official forms which contain machine printed text of different font sizes and various patterns of handwriting are considered for verification and validation of the proposed algorithm. It also includes some graphical components like seal.

Figure 1 shows the architecture of the proposed system.

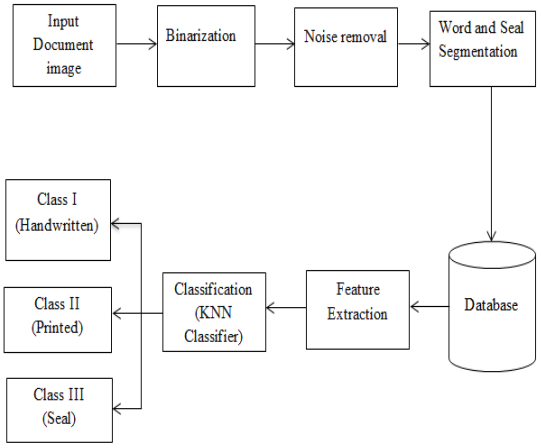
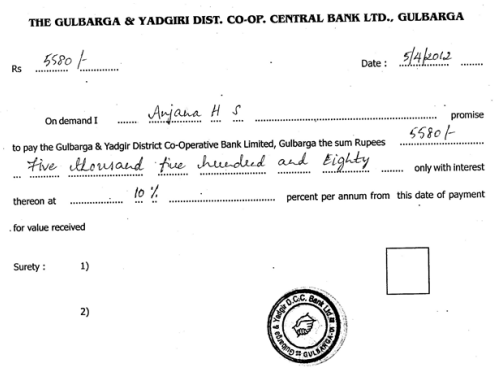


Fig 1: Architecture of proposed system

The input image is scanned and binarized by Otsu's method. The images are filtered by morphological operations to eliminate the noise. Segmentation is performed twice by rule – based method. First to separate the text as handwritten text and printed text and secondly to separate the graphical component from the text component of document image. These separate entities are then stored in the database. The features like perimeter, form factor, density, vertical projection variance etc. are extracted and submitted to the classifier. The KNN classifier is used for classification of the entities of the document image.

Image Database

Somereal images of DCC bank documents are obtained through the HP Photosmart C4388 series scanner machine and remaining images are obtained from IAM database 3.0. The real images are scanned at 300dpi (print resolution) which generates an image of 3510*2550 pixels. The acquired images may be color, greyscale and binary image. The images are in the landscape mode and therefore they are rotated to 90 degree to the right or left. The scanned images are in JPEG format (Joint Photographic Experts Group).


THE GULBARGA & YADGIRI DIST. CO-OP. CENTRAL BANK LTD., GULBARGA

Rs 5580/- Date : 5/4/2012

On demand I Anjana H S promise

to pay the Gulbarga & Yadgir District Co-Operative Bank Limited, Gulbarga the sum Rupees 5580/-

Five thousand five hundred and Eighty only with interest

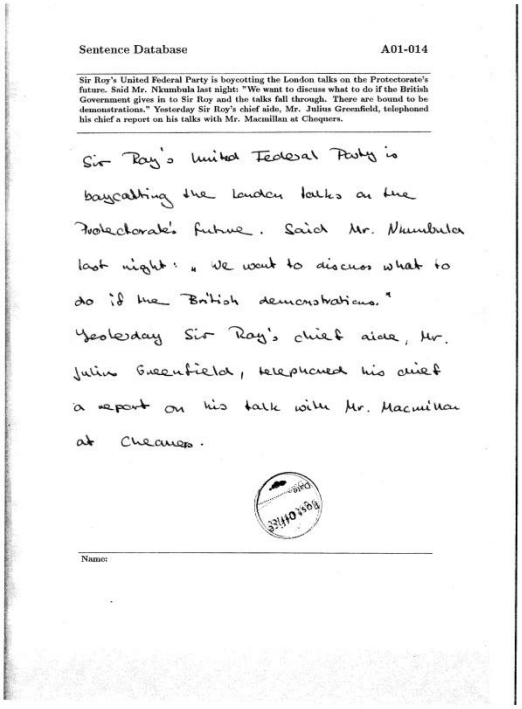
thereon at 10 % percent per annum from this date of payment

for value received

Surety : 1)

2)

Fig 2(a): Sample input image


Sentence Database A01-014

Sir Roy's United Federal Party is boycotting the London talks on the Protectorate's future. Said Mr. Nkumbula last night: "We want to discuss what to do if the British Government gives in to Sir Roy and the talks fall through. There are bound to be demonstrations." Yesterday Sir Roy's chief aide, Mr. Julius Greenfield, telephoned his chief a report on his talks with Mr. Macmillan at Chequers.

Sir Roy's United Federal Party is boycotting the London talks on the Protectorate's future. Said Mr. Nkumbula last night: "We want to discuss what to do if the British demonstrations." Yesterday Sir Roy's chief aide, Mr. Julius Greenfield, telephoned his chief a report on his talks with Mr. Macmillan at Chequers.

Name:

Fig 2(b): Sample Input image of IAM database

Pre-processing

The acquired image may be degraded due to the diversity in the quality of paper, ink, dust and the scanner machine used. Hence, it is essential to perform pre-processing on the sample image. Pre-processing is composed of sequence of steps used to generate an enhanced version of the input image. The rate at which the image is pre-processed affects the accuracy rate of classification and identification. Figure 3 shows the different stages in pre-processing.

Different Stages

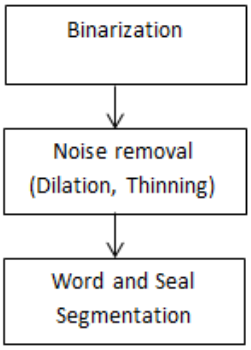


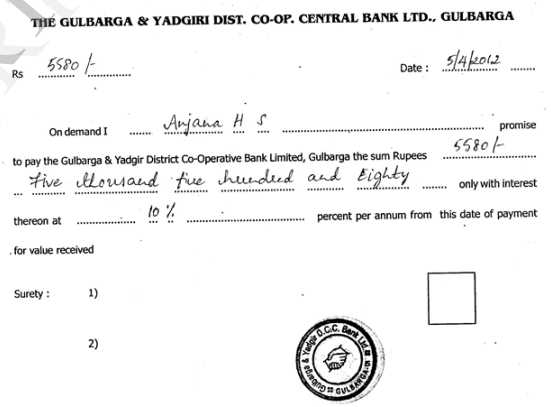
Fig 3: Different stages in pre-processing

The color image is converted to grey scale image. The greyscale image is transformed to binary image by 'Otsu's method'. The color image cannot be directly converted to binary image. Otsu's method is based on a discriminant analysis which is unsupervised and non-parametric method. This method assumes, the image to be thresholded contains two classes of pixels foreground pixels and background pixels. Foreground pixels are represented by 1's and background pixels by 0's. The optimum threshold is calculated for separating the foreground and the background information.

3.1 Word segmentation

Sequence of characters or letters form a word. A single character 'a' and 'I' is also considered as a word in the text. Words are the fundamental units of a document. The linear sequence of words form a line and these lines form a paragraph. There may be any number of paragraphs in the document.

The isolated words and numerals help the system to classify the text. The printed text is uniform in nature whereas handwritten text is non-uniform as it depends on the writing pattern of the writer. Figure 4(a) shows the original input image and figure 4(b) shows the isolated words, numerals and symbols.


THE GULBARGA & YADGIRI DIST. CO-OP. CENTRAL BANK LTD., GULBARGA

Rs 5580/- Date : 5/4/2012

On demand I Anjana H S promise

to pay the Gulbarga & Yadgir District Co-Operative Bank Limited, Gulbarga the sum Rupees 5580/-

Five thousand five hundred and Eighty only with interest

thereon at 10 % percent per annum from this date of payment

for value received

Surety : 1)

2)

Fig 4(a): Original image

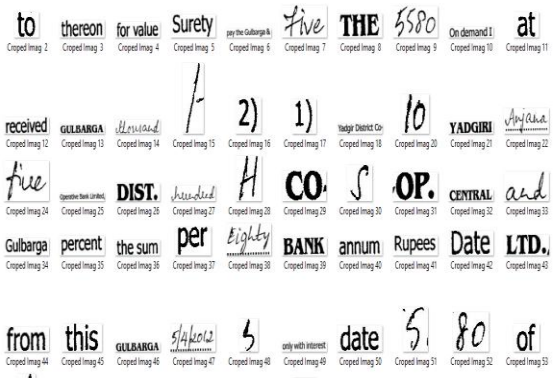


Fig 4(b): Segmented words, numerals and special symbols

Rule based method for Word Segmentation

Rule based system should expose the knowledge hidden in data, providing logical justification for drawing conclusions. Rules are used to support decision making in classification.

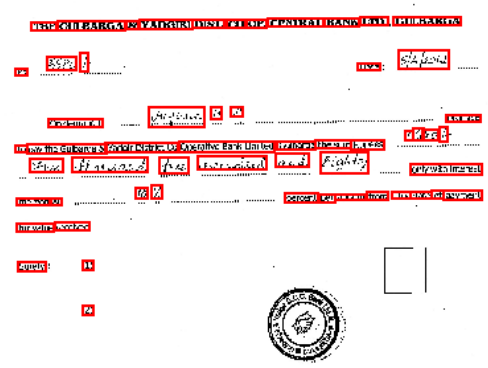


Fig 5: Word segmentation result

The logical rules like if threshold area is greater than the area of other image objects then the index of the image is 1. The threshold area is set to 81. The higher height and lower height is set to 100 and 10 respectively. The lower width is set to 10. These values are obtained by trial and error basis. The text, numerals and special symbols are identified and segmented within the specified range as shown in figure 5.

3.2 Seal Segmentation

Administrative and other official documents contain graphical objects such as seals, logos, stamps, table, maps, engineering drawings etc. Seals are usually seen in official documents, post cards, inland letters, and envelopes.

Seal is a complex entity that has mixture of textual and graphical components which indicate the origin and usage of seal. Some seals consist of variable fields such as date, which may provide the sending or receiving information of a document. A document image may contain any number of seals. The graphical entities are larger than the textual entities with closed connected points. They consist of uniform regions and are highly structured. The text and the seal may or may not be of same color. Seals may be placed in any arbitrary orientation. Detection of this synthetic entity is carried over the entire document which increases the performance of document retrieval. The problem of locating a graphical symbol in a document image is called 'symbol spotting'. When the idea of symbol spotting is extended to a document image database i.e. a digital library, then it is known as 'symbol focussed retrieval'.

Fig 5: Segmented seals

Based on the contour of the graphical image object (seal), the seal is located. Due to imperfect ink condition, some parts of the seal imprint might have different intensity. Figure 5 shows the segmented seal with different intensities and overlapped seals. If the intensity of seal imprint is very less, then it is eliminated as noise in the previous (noise removal) stage. If the desired intensity of seal imprint exists, then segmentation is carried out by performing fundamental morphological operations. Dilation is performed twice by using a line structuring element. The contents of the document image are in the foreground. This foreground content thickens when dilation is performed. The seal is dilated horizontally as well as vertically controlled by a line structuring element. After dilation thinning operation is performed to fill the holes which are created during dilation. Finally, the image is segmented into its individual components. The overlapped seals are also segmented. The segmented seals from all the images are collected and saved in the database. The seal is segregated based on the outer boundary frame, irrespective of the text circumscribed in it.

Rule based method for seal segmentation

The seal has a standard size which does not vary. The rules are predefined for height and width of the seal. The threshold area of the seal is set to 40000 by trial and error basis. The threshold area is always greater than the area of other image objects. The seal is identified and segmented at this particular threshold.

3.3 Connected Component Labelling

A connected component in a binary image is a set of pixels that form a connected group. For example, the binary image below has three connected components. Connected component labelling is the process of identifying the connected components in an image and assigning each one a unique label as shown in the figure 7.

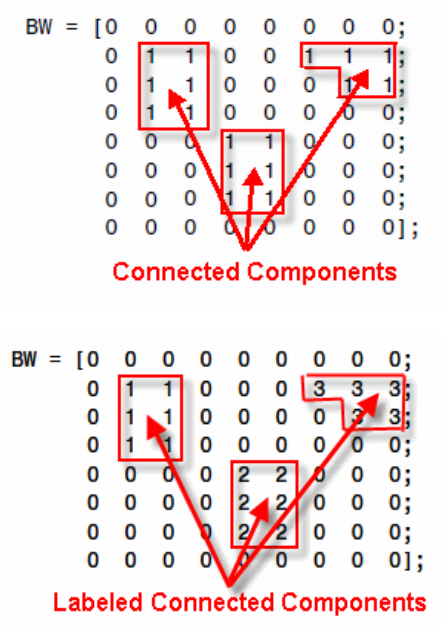


Fig 7: Connected components and labelled connected components

The pixels labelled 0 are the background pixels and the pixels labelled 1 are the foreground pixels. In the figure the pixels labelled 1 is the first object, the pixels labelled 2 is the second object and so on. After assigning the label, a label matrix is generated. The input image is of the unsigned integer and non-sparse. The output is the binary image which is logical. All white pixels are represented by 1 and black pixels by 0. The neighbourhood specifies the type of connectivity. Here the connectivity is 8. The number of connected objects, the size of the image and the number of pixels belonging to each connected component are identified. The neighbourhood which is connected is symmetrical about its pixel that is at the centre. The label matrix is built to visualize the connected components. The size of the label matrix depends on the size of input image and the structure of the connected components. The first object is made up of pixels labelled 1, the second object is made up of pixels labelled 2 and this process is continued until all the objects are labelled which form the connected components. Then for each connected component a bounding box is formed.

The bounding box of a connected component or symbol is defined to be the smallest rectangle which circumscribes the connected component or symbol. A bounding box can be represented by the x, y co-ordinates, width and height. Each bounding box is considered as a smallest entity on the page. It is less computationally intensive. The number of connected objects may or may not be equal to the number of bounding boxes in the image document.

3.4 Feature extraction and selection

The features are designed to differentiate between English printed text, handwritten text and seal which are the distinct components of a document image. These components along with the noise have different visual appearance and physical structures. The potential features of document image components are collected for training. The number of training samples in the database is directly proportional to the accuracy rate of the system.

The feature extraction has an impact on the efficiency of the classification and identification system. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately and captures the different characteristics of the document.

The shape of the connected component is an important feature for classifying the handwritten and printed text and also seal. Based on the shape of the connected component, the features are extracted which aids in classifying the text as printed text and handwritten text and graphical entity. The connected component is referred with its neighbourhood surrounding as context. In document images, most of the text components are smaller than the graphical or non-text components. Therefore size information also plays a significant role in separating the text and non-text component. In addition to size, other features are also considered.

The features that are extracted from the whole document image are the global features. The features which are extracted from the blocks identified during segmentation or from subdivision of the document are local features.

Based on the shape of connected component of a word and seal the following local features are extracted:

1. Area: Area is defined as the actual number of white pixels in the region.

2. Perimeter $= 2 \times breadth + 2 \times height$

3. Form factor $= \dfrac{4 \times pi \times Area}{Perimeter^{2}}$

4. Major Axis: It is defined as the length of the major axis of the ellipse that has the same normalized second central moments as the region.

5. Minor Axis: It is defined as the length of the minor axis of the ellipse that has the same

normalized second central moments as the region.

6. Roundness $= \frac{4 \times Area}{pi \times MajorAxis^2}$

7. Compactness $= \frac{\sqrt{(4 \times Area)/pi}}{MajorAxis}$

8. Density: This is given by

$$Density = \frac{Area\ of\ white\ pixels\ within\ BB}{Area\ of\ bounding\ box\ (BB)}$$

9. White pixels of each line (WPEL): This is given by

$$WPEL = \sum \frac{Number\ of\ white\ pixels\ of\ each\ line}{Width\ of\ BB}$$

10. Vertical projection variance: The vertical projection of white pixels within the bounding box is found and then the variance of only the vertical coordinates of the vertical projection profile is computed.

A feature vector is derived from the mean of above ten features. The text is identified as handwritten and printed text using the above features and based on below conditions:

1. If white Pixel density greater than 70% then text is classified as the printed text
2. If the mean of white Pixels in each line is greater than 75%, it is classified as printed text
3. If Vertical projection Variance ≤ 99 then text is classified as the handwritten text

The seal is identified based on three significant features namely area, major axis and roundness. Some of the objects in the image may have same size and area. Therefore based on the below condition, seal is classified:

If roundness is greater than 80% then the object is classified as a seal

The set of features are computed for the input image to the final stage of recognition. The features would carry sufficient information from the input data to perform machine learning and pattern recognition tasks accurately.

3.5 Classification

The KNN classifier is used for classification. The k nearest neighbours is determined based on the Euclidean distance and is obtained as

$$d(x,y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

The best choice of k depends on the data. Here the value ofk is 3 which is constant. K is the number of classes to be classified. Ties also arise when two distance values are same. In order to avoid such situation, a random component is added. The distances are sorted in the ascending order.

The below algorithm gives a sequence of steps performed in the system

Algorithm: Classification of Document Image Components

Input: Scanned document image

Output: Classification of Image Components in a document

Start

Step 1: Image acquisition

Step 2: Image conversion to binary format by Otsu's method

Step 3: Perform morphological operations such as dilation and thinning to eliminate noise

Step 4: Segment the document image by Rule based method

Step 5: Label each segment of the document image through CCL

Step 6: Form the bounding box for the document image components

Step 7: Generate a feature vector for all the connected components

Step 8: Based on the extracted shape features classify the image components using KNN classifier with k=3.

Stop

4. Experimental Results

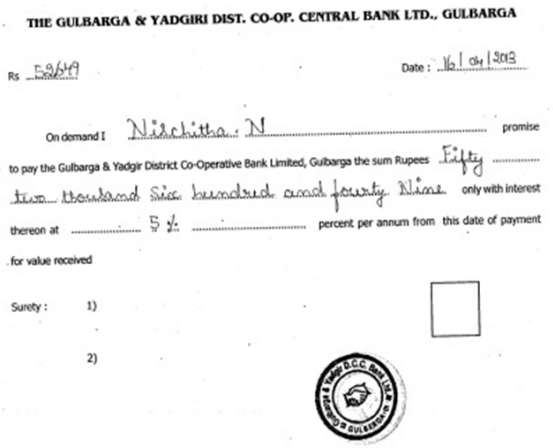
Dataset

We have collected two different official documents of the same bank. The blank documents were filled by different professionals such as graduates, post–graduates, doctors, business people, different institutions, organizations etc. These documents are noisy which consist of printed text, handwritten text and a seal which is a graphical entity.

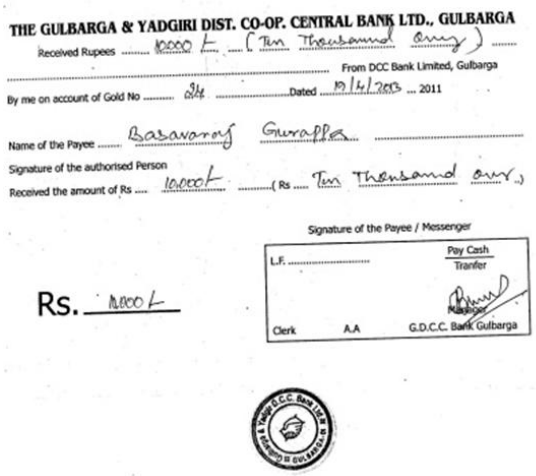
The proposed system is tested on a sample of elevendocument images collected from DCC bank and five images from IAM database 3.0. A total of 1304 segmented images are used for training which includes 556 handwritten images, 728 printed images and 20 seal images. Table 1 gives the detailed information of number of trained images.

Table 1: Statistical Information

Dataset	Features Considered	Segmented Training Samples
Handwritten text	10	556
Printed text	10	728
Circular Seal	3	20
Total		1304



(a)



(b)

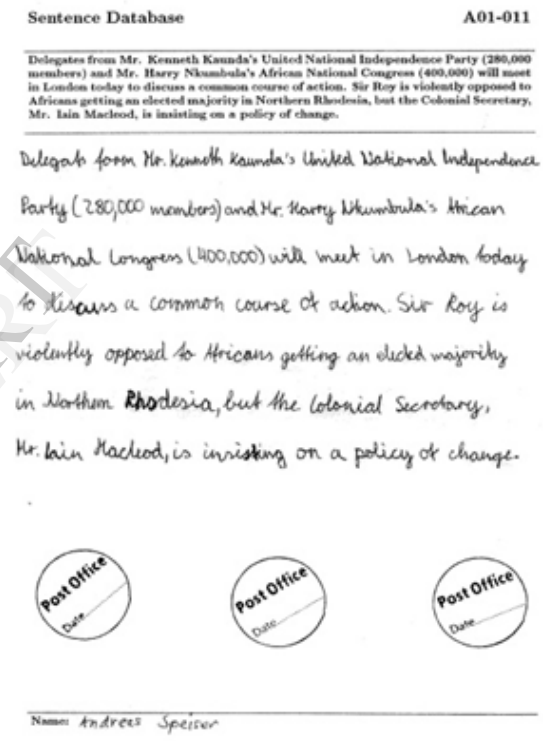
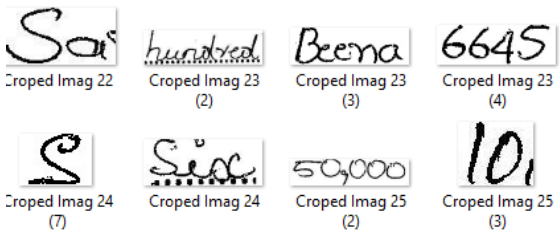


Fig 8 a, b and c: Image samples used for testing



(a)

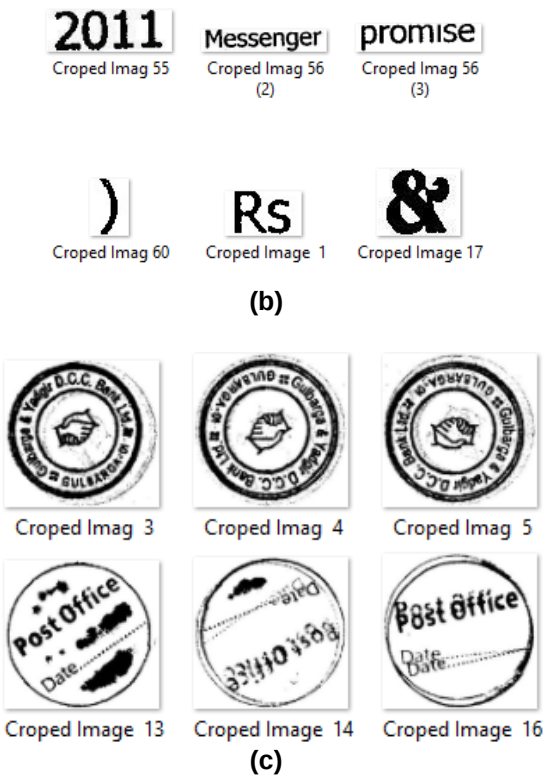


Fig 9:(a) Samples of handwritten training images (b) Samples of printed training images (c) Seal samples used for training

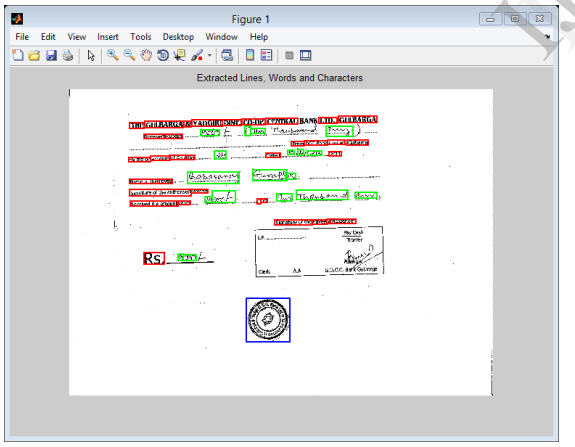


Fig 10(a): Output image of bank

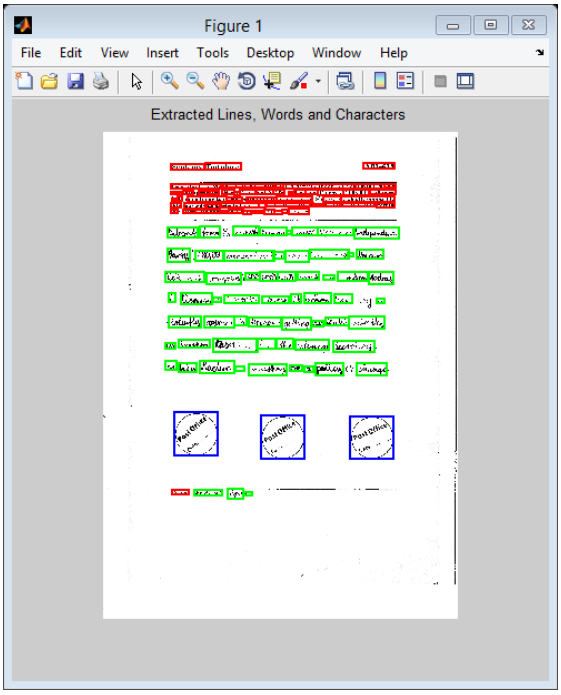


Fig 10(b): Output image of IAM database

The red color bounding box indicates printed text, green color indicates handwritten text and blue color signifies the seal in the figure. Figure 11 shows the overall accuracy rate of the image components.

Evaluation metrics

The overall performance of the system can be evaluated as given by the following equations:

$$\text{Acc. rate (handwritten text)} = \frac{\textit{Total correctly classified HW}}{\textit{Total HW}} \times 100$$

HW- Handwritten words

$$\text{Acc. rate (printed text)} = \frac{\textit{Total correctly classified PW}}{\textit{Total PW}} \times 100$$

PW- Printed words

$$\text{Acc. rate (Seal)} = \frac{\textit{Total correctly detected seals}}{\textit{Total seals}} \times 100$$

Table 2 gives the accuracy rate of test images considered for the experimental result and the overall accuracy is 91.057%.

Table 2: Overall classification rate in different document images

Document images	Accuracy rate of image components (%)
Scan0001	90.285
Scan0002	85.411
Scan0003	88.5
Scan0004	88.444
Scan0005	84.222
Scan0006	92
Scan0007	92
Scan0008	93.904
Scan0009	90.666
Scan0010	82.317
Scan0011	92.666
Scan0012	94.846
Scan0013	92.203
Scan0014	99.346
Scan0015	97.366
Scan0016	92.740
Average	**91.057**

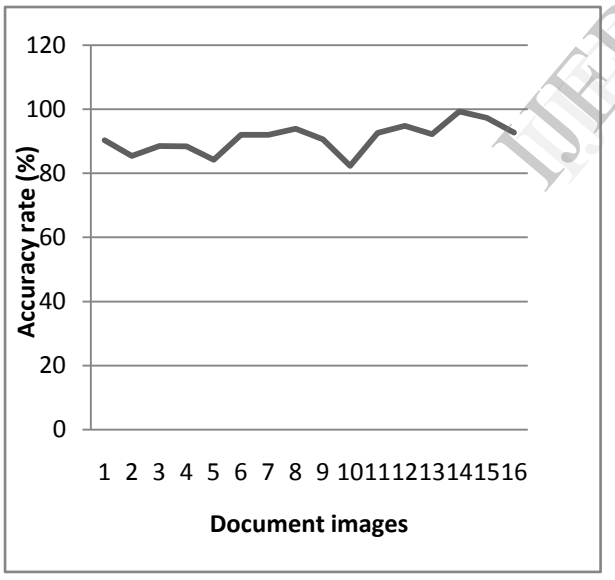


Fig 11:Overall accuracy rate of the image

5. Conclusion

The processing of complex documents is a difficult task for the purpose of classifying the distinct image entities in an automated approach. In the proposed approach, a new system which classifies three distinct image objects such as handwritten text, printed text and seal is introduced. The classification accuracy mainly depends on the training samples and rate of segmentation. The user has to select only the query image to label the connected components and extract the features. The KNN classifier is used to classify the image objects into their respective classes as handwritten text, printed text and seal. Experimental results proved the effectiveness of the proposed approach which is 91.057% on an average.

In future we can classify more than three image objects and also identify individual objects to know the ownership of the document. The seal imprint which has less intensity can also be identified in future which increases the overall classification rate. The text lines and other graphical entities can also be identified and segregated in future to enhance the efficiency of the proposed system. Therefore the classification of different entities of a document image is necessary with the official and government documents.

References

[1] R.Ajay, J.Ayoob, K.Manoj, "Document Image Analysis". IJARCSSE Vol. 2, May 2012
[2] U.Patil, M Begum, "Word Level Handwritten and Printed Text Separation Based on Shape Features". IJETAE Vol. 2 April 2012
[3] Konstantinos Zagoris, Ioannis P, Apostolos A, "Hanndwritten and Machine Printed Text Separation in Document Images using the Bag of Visual Words Paradigm", Int Conf. on Frontiers in Handwriting Recognition 2012
[4] T Kasar, A.G.Ramakrishnan, Amey D, Abhishek Sharma, "Text Extraction using Color-based Connected Component Labeling", Cent Conf Electrical Engg, Indian Institute of Science Bangalore 2011
[5] Dhore M P, Thakare.V.M, Kale.K.V, "Morphological Segmentation in Document image Analysis for Text Document Images". IJCIT Vol. 2, Oct 2011
[6] M.S.Shirdhonkar, Manesh.B.K, "Discrimination between Printed and Handwritten Text in Documents". IJCA Spl Issue on RTIPPR 2010
[7] P P Roy, U Pal, Josep Llados, "Document Seal Detection using Ght and Character Proximity Graphs", Computer Vision Centre Dec 2010
[8] Shazia A, Dr Mehraj-Ud-Din and A Quyoum "Document Image Processing – A Review" IJCA Vol. 10, Nov 2010
[9] S S Bukhari, M.Ibrahim.A and F.Shafait "Document Image Segmentation using Discriminative Learning over Connected Components". ACM June 2010
[10] M.Benjelil, Slim Kanoun, Remy, Adel M "Steerable pyramid based Complex Document Image Segmentation". In IEEE editor ICDAR 2009
[11] Lincoln Feria and Angel Sanchez, "Automatic discrimination between printed and handwritten text in documents". Brazilian Symposium on Computer graphics and image processing
[12] Chen Huang, Sargur N Srihari, "Word Segmentation of Offline Handwritten Documents", 2008

[13] D Keysers, F.Shafait and T.M.Breuel, "Document Image Zone Classification – a simple high-performance approach". Int. Conf. Computer Vision Theory and Applications Mar 2007
[14] R.Kasturi, Lawrence and O'Gorman "Document Image Analysis: A primer". Sadhana Vol. 27 Feb 2002
[15] Jayant.K, Jaishanker.P, David.D, "Document Image Classification and Labeling using Multiple Instance Learning"
[16] Olivier.A, "Document Image Recognition and Classification"
[17] Santanu.C, Megha.J, Sumantra.D.R, "Model-Guided Segmentation and LayoutLabelling of Document Images using a Hierarchical Conditional Random Field"
[18] Claude Faure, Nicorevincent, "Document Image Analysis for Active Reading", International Workshop on Semantically Aware Document Processing and Indexing, 2007
[19] N.Chen, "A survey of document image classification: problem statement,Classifier architecture and performance evaluation", August 2006
[20] Y.Zheng, David.D, "Machine Printed Text and Handwriting Identification in Noisy Document Images", IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 26 March 2004
[21] Henry.S.B, "Difficult and Urgent Open Problems in Document Image Analysis for Libraries", Proceedings of first International Workshop on Document Image Analysis for Libraries 2004
[22] Kathrin.B, Edward.L, "Resolution-sensitive Document Image Analysis for Document Repurposing"
[23] Bing-Fei Wu, Yen-Lin Chen, Chung-Cheng, and Chorng-Yann Su, "A Novel Image Segmentation Method for complex document images" CVGIP 2003
[24] Y Zheng, H Li and D Doermann, "The Segmentation and Identification of Handwriting in Noisy Document Images", Proc. Int Workshop Document Analysis Systems , pp 95-105 2002
[25] Daniel.X.Le, George R Thoma, and Harry Wechsler**, "**Automated Page Orientation and Skew Angle Detection for Binary Document Images"
[26] Daniel.X.Le, George R Thoma, "Document Skew Angle Detection Algorithm", April 1993
[27] D.S.Bloomberg, "Multiresolution Morphological Approach to Document Image Analysis" October 1991
[28] Sivaramakrishnan. R, Arun.C, "Connected Component Labeling and Extraction Based Interphase Removal from Chromosome Images", International Journal of Bio-Science and Bio-Technology, Feb 2013
[29] Rafael C Gonzalez, Richard E Woods, "Digital Image Processing", Pearson Education, Third Edition
[30] Rafael C Gonzalez, Richard E Woods, Steven L Eddins, "Digital Image Processing", Pearson Education, Third Edition
[31]http://www.iam.unibe.ch/fki/databases/iam-handwriting-database