DOCUMENT CLASSIFICATION: A TECHNICAL REVIEW

Mr. Manish Vala

Abstract : Document classification is used for identify the proprietary of complex document. Identify proprietary of any document is very difficult stuff in the area of image processing. There are various way to identify the proprietary of document like, based on signature, logo, seal and many more. We have searched for document classification based on logo and seal, and we surveyed literature related to our work. Majority authors used texture feature extraction author used Discrete Wavelet Transformation (DWT) and Fast Fourier Transform (FFT) for feature extraction and for classification they were used Know Nearest Neighbor (KNN), Neural Network (NN) and support vector machine (SVM).

Keywords: DWT, FFT KNN, NN, SVM

I. INTRODUCTION:

Digital Image processing is the wide and interesting research area due to progress in general electronic technology and document image processing for variety of applications. "Complex documents are those documents containing printed or hand written text or numerals, graphical objects like seal, logo, diagrams, charts, tables, graphs and many more."[3]Nowadays, every organization is moving towards a paperless automation, but before a decade it was basically printed and hand written documents were used. Past document were difficult to analyse due to quality or other factor and difficult to organize documents in well manner. It is motivation factor for processing and analysis image of document.

The need of classification of complex document is to digitalize the current document for fast access and to retrieve it in future. Majority of researcher has used tobacco-800 dataset for classification of complex document. It is widely used in different applications like banking sectors, government organization, private organization and universities. By combining many of the methods it is concluded that KNN classifier provides greatest accuracy that is 87.09%.

In this, paper we were reviewed about the methods and classification based on document classification and we have done the comparison of different classifier and methods. It is used for difficult to organize document in well manner, and possibility of degradation of past document. It can also unable to identify the number of document issued by any organization. Section 2 describe the literature review, Section 3 describes technical details of all the methods and conclusion of complex document classification is summarized in section 4.

Our aim for reeving this paper is to compare all features extraction method and all classifier for classified the document. And from this review we proposed our own model which will help to classify the document.

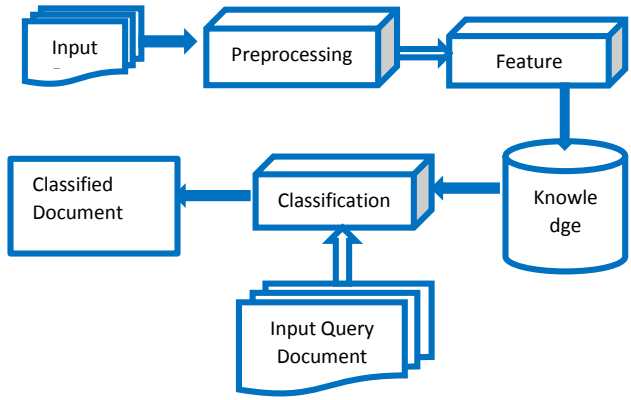

Figure 1 Block Diagram of the Proposed System

The proposed system addresses the problems of logo, printed/handwritten texts and seal identification and document classification using potential features and suitable classifiers to enable to obtain the high accuracy.

II. LITERATURE REVIEW:

Many research has been carried out in last decades on document classification. In above work mainly focused on character, word, line and paragraph. In [1, 2] author worked on printed and handwritten characters, bilingual and multilingual scripts.

Large amount of work has been carried out on complex document containing logos. Automatic logo detection and recognition continues to be great interest to the document retrieval community as it enable effective identification of the source of a document. In this, a new approach is used for detection of logo and extraction from the document image that robustly classifies and precisely localize logos using a boosting strategy across multiple image scales. In this three phases pre-processing, feature extraction and feature matching techniques are used[3].

S. Soma [4] proposed algorithm to recognize logo which used DWT and FFT texture features with NN, KNN and SVM and get average recognition accuracy as 67.74%, 79.35% and 87.09% respectively. In figure 1, the input image is required for both training and testing phase and pre-processing is perform using colour. Translation and morphological operation. Features are extract using Discrete Wavelet Transform (DWT) and Fast Fourier Transform (FFT). Trained vector

generated after extracting the features and stored as knowledge base dataset. The best feature is inputted to the classifiers and compared with qualified data to gain the standard logo.

The document image analysis can be categorized as 'Textural processing' and 'Graphical processing' where the region of interest will be image objects like graphs, picture, stamps and etc. The segmentation image is also one of the challenging research area. In this Support vector Machine (SVM) classifier with RBF kernel is used for pattern matching. The proposed algorithm is experimented on a data set of letter from Gulbarga University. The average recognition accuracy of this algorithm is 85.42%.In below figure 2, author explain stages of Pre-processing. Pre-processing is collected order of steps used to make an improved version of the input image. The rate of which the image is pre-processing affects the correctness rate of categorization and identification. Figure 3 Pre-processing always common with this phase. Binarization means conversion of color and grayscale image to binary format. The color image is change to grayscale image but the color image can't be easily changed to binary image. 'Otsu's using this method grayscale image easily transformed to binary image [5].

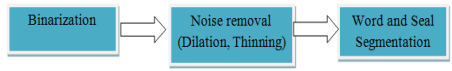

Figure2 Stages of Pre-processing [6]

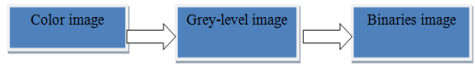

Figure 3 Process of Binarization [6]

In this paper, researcher has used morphological dilation operation to merge separated parts of logos to improve logo detection performance.Tobacoo-800 dataset use for composed of 1290 document images,416 document images occupy logos. This method provides75% accuracy in vertical detection and 74% accuracy in horizontal logo detection [6].

In this paper, researcher has discussed about types of evaluation performance of logo based document retrial using different parameters. They have not used different classifiers and algorithms also discussed and divide all methods and classifiers based on their usage of application and also compared them. Author has highlighted SURF (Speed Up Robust Features) and SIFT (Scale Invariant Feature Transform) algorithm[7].

In this paper, author proposed a new approach that is fisher classifier and multi scale approach, the accuracy of feature extraction is 84.2% and precision is 73.5%. They used tobacoo-800 data set for logo detection and extraction. Author used accuracy and precision as metrics to evaluate logo overall detection performance[8].

$$\text{Accuracy} = \frac{\text{\# of correctly detected logos}}{\text{\# of logos in groundtruth}}$$

$$\text{Precision} = \frac{\text{\# of correctly detected logos}}{\text{\# of detected logos}}$$

After performing the above two function they achieved different accuracy and precision by different method that given in below table.

Table 1: Summary of logo detection [8]

Description	Accuracy	Precision
Improved spatial density [10]	39.3%	32.1%
Fisher classifier only, i.e. |S| = 1	59.2%	41.7%
Multi-scale approach with |S| = 2	57.0%	68.1%
Multi-scale approach with |S| = 3	84.2%	73.5%

In [9] author update line segment Hausdorff distance was proposed for the logo recognition. The new approach features to incorporate structural and spatial information to compute dissimilarity between two sets of line segment rather than two sets of points.

In this paper, researcher had used the object detection technique and nearest neighbor classifier for Detection of Variable Regions in Complex Document Images. In this paper Tobacco800 dataset and Maryland University's document was used for detection of complex document image. Method for variable region detection in degraded or noisy document images is comprised of four phases. The

document is assumed as input and forwarded for preprocessing in phase 1. Object finding in phase 2. Histogram of grade feature are computed from the object detection in phase 3 and features are confidential using nearest neighbor classifier in phase 4[10].

As we discussed before document contain handwritten and printed text so, many time to identify any document form the text we need to perform the author designed algorithm to recognize the hand written kannda vowels based on shape features such as normalized chain codes and wavelet filters. They took size 40*40 and KNN classifier is applied for classification of vowel. The database for vowels consists of 1400 samples of the handwritten vowel images, with 100 images in place of each class/vowel. For result calculation author used 2-fold cross confirmation technique. The KNN classifier classifies test vowels to a class based on K nearest neighbor. The experiments be carried out by varying the values of K i.e. K=1, 3, 5 and found best result when K =3. The average identification rate was obtained for Kannada vowels is 95.07 that we described in Table 2[2].

Table 2 Experiment Comparative Analysis of Chain code and Wavelet Filter.

No	Method	Features dimension	Recognition Accuracy in%
1	Chain code	08	77.00
2	Normalized Chain code	08	80.00
3	db filter	14	89.00
4	Chain code and Wavelet filters	22	92.15
5	Normalized Chain code and Wavelet filters	22	95.07

In [11] author discussed about the multi-level staged approach to logo recognition. They used global invariants to prune the database and local affine invariants to obtain a more refined match. They took an invariant signature features which can be used for matching under a variety of transformations. They also computed Euclidean invariants, and show how to extend them to capture similarity, affine

and projective invariants when necessary. Authors worked on feature detection, feature extraction and local invariant algorithms and also successfully demonstrate the same on a small database.

III. TECHNICAL DETAILS OF METHODS:

I. **Neural Network:** Image processing with NN can helps in many areas like defense department, automatic and transport, industrial surveillance and many more. By using NN we can improve the result of classification, identification, and authentication. With NN we can resolve the issues of object matching and object recognition [12]. In [4] author worked on document classification with NN classifier and got 67.7 accuracy.

II. **Support Vector Machine:** Shidevi Soma used the Support Vector Machine classifier with RBF kernel is used for the pattern matching on the letter of gulbarga university and achieved 85.42% accuracy [5].

III. **Know Nearest Neighbor:** In [4] author worked on document classification with KNN classifier and got 79.35% accuracy. In [6] KNN classifier is used for the logo recognition, author implement KNN and got 100% accuracy in without variation document and got 92.5% accuracy in variation document.

IV. CONCLUSION:

During literature review we have studied the different feature extraction techniques as well as discussed about different classifier like NN, SVM and KNN. In this paper we discussed these techniques for document classification based on texture and graphical object. Identification of isolated object and two level of classification form the blank document, printed and handwritten has been done. From the above literature review we conclude that Support Vector Machine classifier with shape features performed better compared to other features and classifiers with a maximum accuracy varying form 80% to 95%. Still there are many scope improvement in the accuracy and precision of document identification with SVM and other classier.

V. REFERENCES:

[1] B.V.Dhandra, Mallikarjun Hangarge, “On Seperation of English Numerals from Multilingual Document Images”, International Journal of Multimedia (JM), Vol.2, No.6 Nov. 2007, Academy Publisher, Oulu, Finland, pp. 26-33, ISSN: 1796-2048.

[2] B.V.Dhandra, Shashikala P, Gururaj M, "Kannada Handwritten Vowels Recognition Based on Normalized Chain Code and wavelet Filters", National Conference on Recent Advances in Information Technology, 2014.

[3] Shridevi Soma1, B.V Dhandra2, "Automatic Logo Recognition System from the Complex Document using Shape and Moment Invarient Features", International Journal of Advances in Computer Science and Technology, Vol 4 No.2, Page: 06-13, 2015.

[4] B.V. Dhandra, Shridevi Soma, Gururaj Mukarambi " Identification of Institutional Logo based on Wavelet Features", Dept. of Computer Science Central University of Karnataka Gulbarga International Journal of Computer Applications (0975 – 8887) Volume 107 – No 15, December 2014.

[5] Shridevi Soma, B.V Dhandra, "A Shape feature based Identification of a Complex Document", National Conference on Digital Image and Signal Processing (NCDISP – 2015), MAEER's Arts, Commerce and Science College, Pune, India, 12th and 13th Feb 2015.

[6] Sina Hassanzadeh and Hossein Pourghassem, "A Novel Logo Detection and Recognition Framework for Separated Part Logos in Document Images" Australian Journal of Basic and Applied Sciences, 5(9): 936-946, 2011 ISSN 1991-8178, Department of Electrical Engineering, Najafabad Branch, Islamic Azad University, Isfahan, Iran.

[7] Raveendra. K, Dr. P V N Reddy and Dr. P V V Kishore, "A Review on Classification and Comparison of Automatic Logo Based Document Image Retrieval Methods and other Applications". International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 24 (2017).

[8] Zhu, Guangyu, and David Doermann, "Automatic document logo detection." Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on. Vol. 2. IEEE, 2007.

[9] Chen, Jingying, Maylor K. Leung, and Yongsheng Gao. "Noisy logo recognition using line segment Hausdorff distance." Pattern recognition 36.4 (2003): 943-955.

[10] Sreelakshmi U.K, Akash V.G and N. Shobha Rani, "Detection of Variable Regions in Complex Document Images", International Conference on Communication and Signal Processing, April 6-8, 2017, India.

[11] David S. Doermann, Ehud Rivlin and Isaac Weiss, "Logo Recognition Using Geometric Invariants", 0-8186-4960-7/93 $3.00 0 1993 IEEE.

[12] Alexandrina-Elena Pandelea*, Mihai Budescu and Gabriela Covatariu, "Image processing using artificial neural", Buletinul Institutului politehnic din iaşi public at de universitatea tehnică, gheorghe asachi" din iaşi tomul lxi (lxv), fasc. 4, 2011

AUTHORS' PROFILE

Mr.Manish Vala is a faculty at Babu Madhav Institute of Information Technology, Uka Tarsadia University. He has more than 6 years of teaching experience in the field of computer science. His interested research area includes Image processing. Currently he is working on Document Classification based on object.