ISSN 2320 - 2602

International Journal of Advances in Computer Science and Technology (IJACST), Vol. 4 No.2, Pages : 06 – 13 (2015)
Special Issue of ICCSIT 2015 - Held on February 9, 2015, Singapore
http://warse.org/pdfs/2015/iccsit2015sp02.pdf

Automatic Logo Recognition System from The Complex Document Using Shape and Moment Invariant Features

Shridevi Soma[1], B.V Dhandra[2]

[1]Dept. of Comp. Sc. & Engg., P.D.A College of Engineering, Gulbarga, Karnataka, India,
shridevisoma@gmail.com

[2]Dept. of Comp. Sc., Gulbarga University, Gulbarga, Karnataka, India,

dhandra_b_v@yahoo.co.in

Abstract

One of the strongest clues for retrieval of content information from complex document images is logo. However, due to the wide range of transformations that an object might undergo, this is also the most difficult one to handle. Logo retrieval is one of the major barriers now a days for image databases being commonly used. Automatic logo detection and recognition continues to be of great interest to the document retrieval community as it enables effective identification of the source of a document. In this paper, a new approach is proposed for detection of logo and extraction from the document images that robustly classifies and precisely localizes logos using a boosting strategy across multiple image scales**.** At a coarse scale, logo recognition system also comprises of three phases: Preprocessing, feature extraction and features matching. For feature extraction 10 shape features and 07 hu's moment invariant features have been adopted. The Euclidian Distance (ED) is taken as a similarity measure parameter for the features matching with Nearest Neighbor, K Nearest Neighbor and Support Vector Machine Classifier are also compared. The accuracy of 96.24% from the SVM classier proved to be potential classifier from the experimental results as compared to 88.21% for NN and 91.46% from KNN classifiers.

Keywords: pre-processing, segmentation, feature extraction, NN, KNN, SVM.

INTRODUCTION

Logos are commonly used in business and government documents as a declaration of document source and ownership. The problem of logo detection and recognition is of great interest in the document domain as it enables us to identify the source of documents based on the organization where a document originates. Facing continually increasing volumes of documents, logo recognition has evolved as a practical and reliable supplement to the recognition of printed text using OCR and analysis of the textual content through natural language processing. In the context of document image retrieval, logos provide an important form of indexing that enables effective exploration of data. Given a large collection of documents, searching for a specific logo is a highly effective way of retrieving documents from the associated organization.

The ability to robustly detect logos and extract them intact from volumes of documents is pivotal for logo recognition. Large intra-class variations of logos and the diverse quality and degradations in captured document images make logo detection a difficult problem. Complicating matters, the foreground content of documents generally includes a mixture of machine printed text, diagrams, tables and other elements.

International Journal of Advances in Computer Science and Technology (IJACST), Vol. 4 No.2, Pages : 06 – 13 (2015)
Special Issue of ICCSIT 2015 - Held on February 9, 2015, Singapore
http://warse.org/pdfs/2015/iccsit2015sp02.pdf

Logo of Philip Morris Management group | Logo of American Tobacco Company | Logo of Gulbarga University

Fig 1: Sample logos of different institutions

From the application perspective, the accurate localization needed for logo recognition poses another major challenge. The logo detection module must consistently output complete logos while attempting to minimize the false alarm rate. To our best knowledge, these key aspects of automatic logo detection have not been addressed in the literature, a formulation of the logo detection and recognition problem that jointly considers logo detection and extraction in a unified framework, an evaluation metric that quantitatively measures the quality of extracted logos, and testing on large document collections that demonstrates achievable generalization performance**.**

Logos are used by organizations to identify themselves on documents. When scanned paper documents are analyzed to produce structural electronic representations or for the purpose of sorting by corporate identity, logo recognition becomes an important component. Successful recognition of logos facilitates automatic source classification of document images and may also be used to determine how best to process the information contained within these particular documents.

Sample logos of different organizations and institutions is presented in Fig 1 above. In the proposed work complex documents images are given as a query image and logo object is extracted using area based segmentation method and matched against a logo database. The results of the logo recognition are used to sort the documents according to the company logo on each. Such an application is useful wherever a large collection of business documents requires sorting and searching into organizational classes.

There are a number of challenges associated with logo recognition. One such challenge includes robustness to noise since many digitized documents containing logos have some degree of noise. A primary source of noise issues may be due to the printing, scanning and faxing of paper documents. Another source of noise is introduced from environmental factors such as smudging of ink, damaged documents and dirty documents. The challenge with logo recognition is accurate logo segmentation. A logo image may not be segmented accurately by detection and segmentation process due to a skew and non-normalized logo images may not be available. These extracted logo images may also contain surrounding text and background artifacts from the original document. The type of logos available also varies quite considerably with types such as line style logos, dense graphical logos, textual logos and a combination of these. With these challenges in mind, a logo recognition algorithm must be robust to noise, slight skew angles, scale and logo structure type. The proposed logo recognition system can be used to recognize and retrieve the most relevant logos from the database when query image presented to the system.

Rest of the Paper is organized as follows. Chapter 2 presents the Literature Survey on the proposed work, Chapter 3 deals with the phases of Logo Recognition System such as preparation of Image Database, Preprocessing, Segmentation, Feature extraction and Classification. Chapter 4 presents the details of Experimental results, finally Chapter 5 concludes the paper and highlights the scope for future work.

LITERATURE SURVEY

Prior Literature[1, 2, 4, 5, 7, 8, 10, 11] focused almost on Logo Recognition. All these studies reveals that logo identification and segmentation approaches are available. In Many of these past works Logo database on University of Maryland has been used[13, 14]. [4] In 1993 David S. et. Al. have presented a multi-level staged approach to logo recognition using global invariants to prune to the logo database and local affine invariants to obtain a more refined match. [5] Doermann et. Al. have applied algebraic and differential invariants for logo recognition by extracting text primitive shapes from logos using specific feature detection and used global and local geometric invariants for matching. [6] Liwei Wang et.al. proposed a method that helps to identify image objects by finding the Euclidean distance between the train and test set. [7] B.V Dhandra et.al. proposed a method to efficiently recognize the Institutional logos using wavelet features by considering 31 logo classes and have compared the accuracy rate of three classifiers namely NN, KNN, SVM. Their experimental result showed the higher accuracy using SVM. [8]M.S. Shirdhonkar et. al. presented an algorithm to detecting logo in the document image using Discrete Wavelet Transform to compute the spatial density of the window.

LOGO RECOGNITION

From the Fig 2 indicating block diagram it is evident that Euclidean distance and Support Vector Machine as the classification systems are the core for this system. However before employing these classifier models it is essential for both input image under training and testing should undergo the phases such as Pre-processing, Segmentation and Feature extraction. Following section deals with all of these phases.

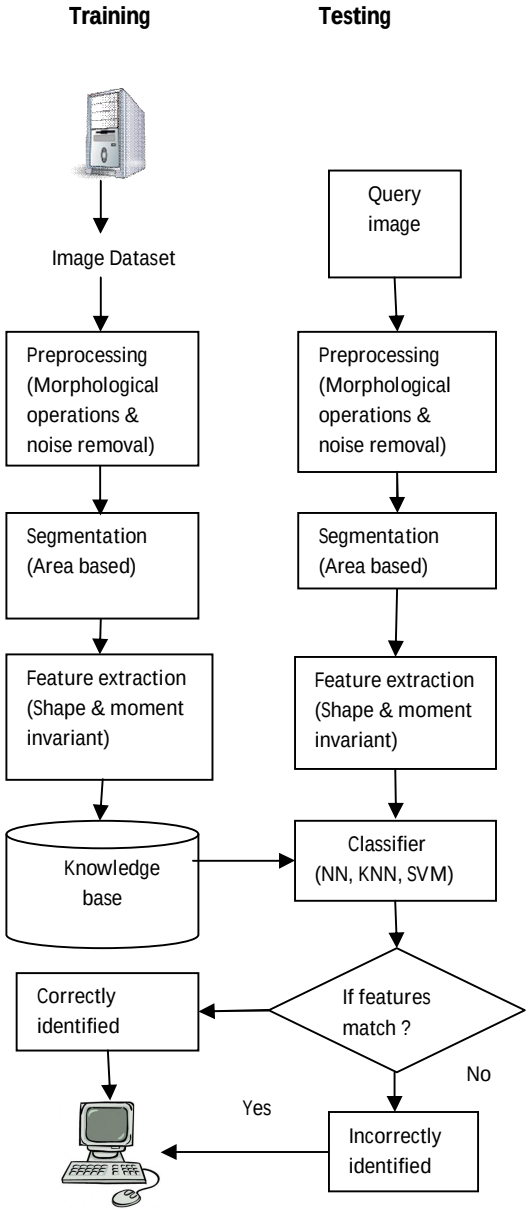

Fig 2 : Block Diagram of the Automatic Logo Recognition

Image Database

Data set for the proposed work contains the images from the Tobacco-800 dataset, a public subset of the IIT CDIP Test Collection[12] constructed by Illinois Institute of Technology. It is a realistic data set assembled from 42 million pages of documents (TIFF images) released by tobacco companies under the Master Settlement Agreement and originally hosted at UCSF [13]. Tobacco-800 contains 1290 document images collected based on a special collection building method[14]. Among these 416 images including logos, In the proposed work 206 images are selected from 416 logo images are used for training set and rest for testing. Also 40 real document images are obtained by scanning them by HP Scanjet G2410 scanner machine with 300 dpi resolution and added to the train set. Total of 246 images are trained and stored as trained set.

Pre-processing

Image pre-processing techniques are necessary, in order to remove the noise and to enhance the quality of the image for better recognition accuracy. Before any image-processing algorithm can be applied on image, preprocessing steps are very important in order to limit the search for abnormalities. The main objective of this process is to improve the quality of the image to make it ready for further processing by removing the unrelated and surplus parts in the back ground of the image. In the proposed work morphological dilation operation is used with line as a structuring element. There are three major steps that are involved in the pre-processing stage.

Binarisation, De-noising and Segmentation. Colour conversion is carried out to get the binary image by Otsu's method in the experiment.

One of the most important problems in image processing is de-noising. Usually the procedure used for de-noising is dependent on the features of the image, aim of processing and also post-processing algorithms. To remove the noise from the input image morphological operations like dilation and filling are used. Morphological operations simplify image data and preserve the essential characteristics of the object shape. These operations are described by a structuring element. Horizontal dilation is

performed. The extent of thickening is controlled by a flat line structuring element which is a set of point coordinates.

Thinning is performed to fill the holes which are unfilled by dilation. This process helps for further analysis of the image.

Image Segmentation

In computer vision, image segmentation is the process of partitioning a digital image into multiple segments (sets of pixels, also known as super pixels). Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. In the similar way for locating the boundaries and edges we take up the watershed algorithm based on rain falling simulation The task performed by the present algorithm is to trace a path for each non-minimum point on the surface (origin) to a minimum (destination), and to mark all pixels along the path with the label of the minimum. This path is a steepest slope line in a lower-complete image. The latter is the transformed gradient image such that any non-minimum pixel has a lower neighboring one. The result is a partition of the image which has the following properties: regions are connected, they do not overlap, and the partition is complete.

A document image may contain text, symbols, numerals, seals, logos, etc. and there might be any number of logos. 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 logo may or may not be of same color and logo is present 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 focused retrieval'.

Based on the contour of the graphical image object, the logo is located. some parts of the logo is degraded. If the logo is noise free, 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 logo 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 segmented logo from all the images are collected and saved in a folder which exists in a database.

Connected Component Labeling

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 labeling is the process of identifying the connected components in an image and assigning each one a unique label as shown in the Fig 3.

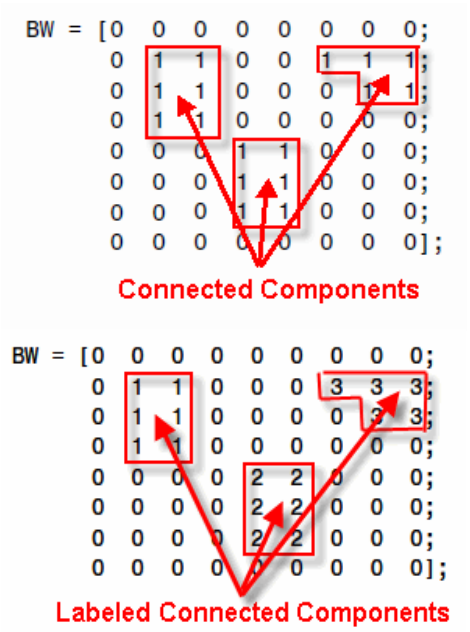

Fig 3: Connected components and Labeled connected components

The pixels labelled 0 are the background pixels and the pixels labeled 1 are the foreground pixels. In the Fig 3 the pixels labeled 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 neighborhood 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 connected neighborhood is symmetrical about its centre element. 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 labeled 1, the second object is made up of pixels labeled 2 and this process is continued until all the objects are labeled 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.

Feature Extraction

When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant then the input data will be transformed into a reduced representation set of features (features vector). Transforming the input data into the set of features is called feature extraction. 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 known as local features.

In the proposed work total of 17 features are used to describe the image, out of which 10 are shapes features and 07 are Hu's moment Invariant features. Based on the shape of connected component of the logo and non-logo objects of complete document image the following local features are extracted:

Shape Features:

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

2. Perimeter: It is defined as the distance around the boundary of the region.

3. Form factor: The pattern of scattering white pixels in an image within the bounding box.

$$\frac{4\pi \times Area}{Perimeter^{2}} \quad (1)$$

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. $$Roounfness = \frac{4 \times Area}{pi \times MajorAxis^{2}} \quad (2)$$

7. Compactness: Compactness is an indication of solidness and convexity. It is given by ratio of the object to the area of a circle with the same perimeter. The maximum value possible is 1 which is for solid circle, image object having complicated boundaries will have lower values. This feature is calculated using Eq.3

$$Compactness = \frac{4\pi \cdot Area_{image}}{Perimeter_{image}} \quad (3)$$

8. Density: Density is defined as the area of white pixels within the bounding box. It is the ratio between area of white pixels within the bounding box and the area of bounding box which is given by:

$$Density = \frac{Area\ of\ white\ pixels\ within\ bounding\ box}{Area\ of\ bounding\ box} \quad (4)$$

9. Mean of black pixel at each line(BPEL) :

$$BPEL = \sum \frac{Number\ of\ Black\ Pixels\ of\ each\ line}{Width\ of\ Bounding\ Box} \quad (5)$$

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.

Moment Invariant Features:

Shape of an object is the characteristic surface configuration as represented by the contour. Shape recognition is one of the modes through which human perception of the environment is executed.. Hu invariants moment are a set of nonlinear functions, which are invariant to translation, scale, and orientation and are defined on normalized geometrical central moments. Hu introduced seven moment invariants based on normalized geometrical central moments up to the third order. Since the higher order moment invariants have resulted higher sensitivity, a set of eight moment invariants limited by order less than or equal to four seems to be proper in most applications. Having normalized geometrical central moments of order four and the lesser ones, seven moment invariants $(\varphi_1 - \varphi_7)$ introduced by Hu and can be computed using equations given below.

$$\varphi_1 = \eta_{02} + \eta_{02} \tag{6}$$

$$\varphi_2 = (\eta_{02} - \eta_{02})^2 + 4\eta_{11}^2 \tag{7}$$

$$\varphi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{8}$$

$$\begin{aligned}\varphi_5 = {} & (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ & + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \\ & [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\end{aligned} \tag{9}$$

$$\begin{aligned}\varphi_6 = {} & (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ & + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})\end{aligned} \tag{10}$$

$$\begin{aligned}\varphi_7 = {} & (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ & + (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03}) \\ & [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\end{aligned} \tag{11}$$

Classification

Classification algorithms typically employ two phases of processing: training and testing. In the initial training phase, characteristic properties of typical image features are isolated and, based on these, a unique description of each classification category, i.e. training class, is created. In the subsequent testing phase, these feature-space partitions are used to classify image features.

A central problem in image recognition and computer vision is determining the distance between images. Considerable efforts have been made to define image distances that provide intuitively reasonable results. Among others, two representative measures are the tangent distance and the generalized Hausdorff distance. Tangent distance is locally invariant with respect to some chosen transformations, and has been widely used in handwritten digit recognition. The generalized Hausdorff distance is not only robust to noise but also allows portions of one image to be compared with another, and has become a standard tool for comparing shapes.

Among all the image distance metrics, Euclidean distance is the most commonly used due to its simplicity. Mathematically Euclidean distance is the "ordinary" distance between two points and is given by Eq.12.

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \tag{12}$$

The Euclidean distance indicates, for each pixel in the objects (or the background) of the originally binary picture, the shortest distance to the nearest pixel in the background (or the objects). A map with negligible errors can be produced.

In the Proposed system Euclidean distance measure is used in all the three classifiers NN, KNN and SVM. For KNN Classifier the value of k depends on the data. Here the value of k is taken as 3. Support Vector Machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. For non linear separable case many kernel mapping functions are used. In this paper a Radial Basis Function is use.

RESULTS AND DISCUSSIONS

A simulation model was developed in MATLAB programming language to implement the system and to analyze its simulation performance. In the quest for finding the best classification procedures. This paper analyzes a machine learning techniques such as Nearest Neighbor, K Nearest Neighbor and SVM classifier and compares the accuracy rate of all of these classifiers and SVM classifier resulted in the optimum identification rate.

The training phase which establishes the lining up of the images is carried out as the current step. Some of the trained images of the logo are shown in Fig 4.

Fig 4: Trained Images

The next step is the testing phase which helps in the extraction of the logo from the document by establishing a series of phases which include Preprocessing, Colour conversion, Obtaining the region of interest of the logo, and finally displaying the extracted logo. This is clearly shown in the Fig 5.

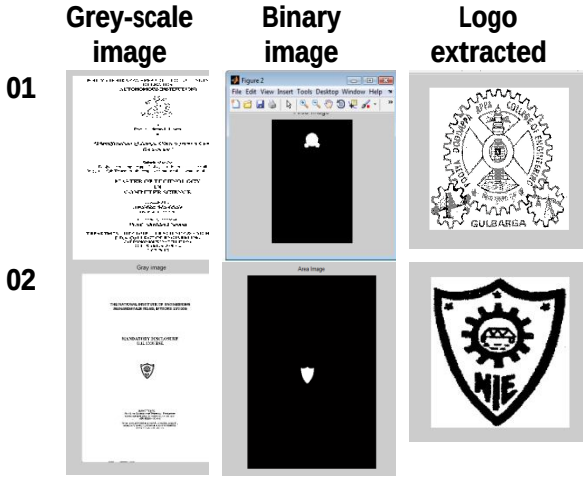

Fig 5: Test Images

The Fig 6 shows the graph of performance of all the three classifier results.

Table – 1: Accuracy rate of NN, KNN and SVM classifiers for logo recognition

Class-ifiers	Number of test images	No. of Correctly identified Logos	No. of Incorrectly identified Logos	Recogniti on rate (in %)
NN	246	217	29	88.2
KNN	246	225	21	91.46
SVM	246	237	9	96.34

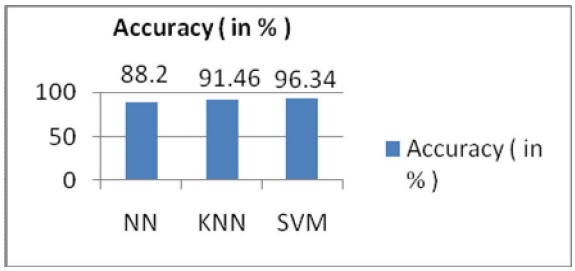

Fig 6 : Graph showing Accuracy of NN, KNN and SVM Classifier

The logo identification system proposed in this work resulted in high accuracy compared to the work carried out previously in reference 7, in which wavelet features are used. The results of proposed work and reference 7 is summarized in Table-2 and graphical analysis is given in Fig-7 below.

Table 2: Performance comparison of proposed system with other scheme

Algorithm	No. of Logo Classes	No. of Features used	Dataset Type	Accuracy NN KNN SVM (in %)		
Logo Identification using Wavelet Features (Ref.7)	31	Wavelet	Scanned Images	67.7	79.3	87.09
Logo Identification using Shape features	48	Shape	Standard Tobacco-800 data set and Scanned images	88.2	91.46	96.34

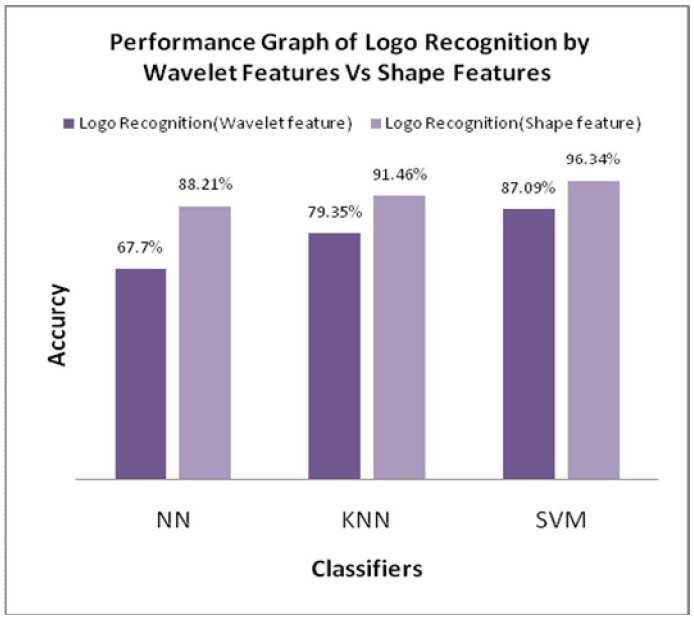

Fig 7: Performance Graph of Logo Recognition by Wavelet Features Vs. Shape Features

CONCLUSION AND FUTURE WORK

The processing of complex documents is a challenging task for the purpose of classifying the distinct image entities in an automated approach. In the proposed work, a system is developed to identify the ownership of a document using logo object present in it. An area based segmentation method is used first to train the logo image and then for testing towards the query image. The combined features from the simple shape descriptors and moment invariant descriptor resulted into the optimal identification compared to the wavelet features for recognition of the logo objects from the document image.

The future scope of the work is to increase the accuracy of the system by combining the image object like logo and name of the ownership of the document image, where name refer to the organisation or institute name.

REFERENCES

[1]. Guangyu Zhu and David Doermann, Automatic Document Logo Detection, In Proc. Int'l Conf. Document Analysis and Recognition, pages 864-868, 2007.
[2]. Omar Mohammed Wahdan, Khairuddin Omar and Mohammad Faidzul Nasrudin, Logo Recognition System Using Angular Radial Transform Descriptors, Journal of Computer Science 7 (9): 1416-1422, 2011 ISSN 1549-3636 2011 Science Publications.
[3]. T. Sedghi, Indexing of Shape Images based on Complementary Composited Features Australian Journal of Basic and Applied Sciences, ISSN 1991-8178. 5(11): 739-743, 2011.
[4]. D.Doermann, E.Rivlin, I.Weiss, Logo recognition using geometric invariants, procee International Conference on Document Analysis and Recognition, page 897-903, 1993.
[5]. D. Doermann, E.Rivlin, I.Weiss, Applying differential invariants for logo recognition, Machine Vision and Application, 9(2), pages 73-86, 1996.
[6]. Liwei Wang, Yan Zhang, Jufu Feng,"On the Euclidean Distance of Images", School of Electronics Engineering and Computer Sciences, Peking University Beijing, 100871, China.
[7]. B.V Dhandra, Shridevi Soma, Gururaj Mukarambi, "Identification of Institutional Logo based on Wavelet Features", Int'l Journal of Computer Applications ISSN 0975-8887. Vol 107 – No.15, Dec. 2014.
[8]. M.S. Shirdhonkar, Manesh Kokare, "Automatic Logo Detection in Document Images
[9]. B.V.Dhandra, Shridevi Soma, Rashmi T, Gururaj M, Classification of Document Image Components, International Journal of Engineering Research and Technology, Vol.2, Issue 10, October 2010, page 1429-1439.
[10]. J.Neumann, K. Samet, and Soffer, Integration of local and global shape analysis for logo classification, Pattern Recognition, 36(12): 3023-3025, 2003.
[11]. Hongye wang, Youbin Chen, Logo Detection in Document Images Based on Boundary Extension of Feature Rectangle, 10th Intl. Conf. on Document Analysis and Recognition, 2009.
[12]. Digital Image using MATLAB by Rafael C. Gonzales, Richard E. Words and Steven L Eddins, Low Price Edition, India.
[13]. Tobacco-800 Complex Document Image Database, http://www.umiacs.umd.edu/zhugy/Tobacoo800.html.
[14]. The Legacy Tobacco Document Library (LTDL), University of California,2007. http://legacy.library.ucsf.edu.
[15]. D. Lewis, G. Agam, S. Argamon, O. Fieder, D. Grossman, and J. heard, Building a test collection for complex document information processing. In Proc. Annual Int. ACM SIGIR Conference, pages 665-666, 2006.