Handwritten Document Image Retrieval

M. S. Shirdhonkar and Manesh B. Kokare

***Abstract*—Many techniques have been reported for handwritten based document image retrieval. This paper proposes a method by using Contourlet Transform (CT) for feature extraction of document images which achieves high retrieval rate. The handwriting of different people is often visually distinctive; we take a global approach based on texture analysis, where each writer's handwriting is regarded as a different texture. The distance measures viz., Canberra distance and Euclidean distances are used as similarity in proposed system. Superiority of Canberra distance is observed over Euclidean distance in term of average retrieval rate. Retrieval results with proposed method are very promising with precisions and recalls.**

***Index Terms*—Document image analysis and retrieval, Contourlet Transform, document similarity measurement, handwritten documents, handwriting classification and retrieval, writer identification.**

I. Introduction

Modern technology has made it possible to produce, process, store, and transmit document images efficiently. In attempt to move towards the paperless office, large quantities of printed or handwritten documents are digitized and stored as images in databases [1]. The popularity and importance of document images as an information source are evident. Many organizations currently use and dependent on document image database. Searching for relevant document from large complex document image repositories is a central problem in document image analysis and retrieval. In this paper, we have proposed automatic handwritten document image retrieval based on writer. Document image retrieval is a very attractive field of research with the continuous growth of interest and increasing security requirements for the development of the modern society. Our objective is to find writer based document image retrieval. We use multi-channel spatial filtering techniques to extract texture features from handwriting images. There are many available filters in multi-channel techniques. In this paper the handwritten document image retrieval is based on the definition of features that can extract over an entire textual handwriting sample. Hence, we propose a global approach of writer based handwritten document image retrieval using contourlet transform. Standard wavelets based tools suffer with their incapacity to locate thin structure of lines and thin variations all along curves. The need for searching scanned handwritten documents are involved in application such as collection dating, works genesis, critical edition, document authentication. In 2003, Srihari et al. [2], have realized the importance of handwritten document retrieval and have presented their retrieval system dedicated to forensics applications such as writer identification. In 2004, Schomaker and Bulacu [3], computes a code book of connected component contours from an independent training set and employs the probability density function of the unknown writing to identify its author. In 2005, Bensefia et al [4], uses local features based on graphemes that are produced by a segmentation algorithm based on the analysis of the minima of the upper contour. In 2006, Pareti and Vincent [5], models the distribution of patterns occurring in handwriting texts by zipt law, the respective zipt curve charactering the writer. In 2007, G.Joutel et al [6], proposed curvelets based queries for CBIR applications in handwriting collections, in this method curvelet coefficients are used as representation tool for handwriting when searching in large manuscripts databases by finding similar handwritten samples. In 2008, Siddiqi and Nicole [7], proposed effective method for writer identification in handwritten documents.This method is based on identify the writing style of an author and then extracting the forms that a writer used frequently. In 2009, Siddiqi and Nicole [8], proposed a set of chain code based features for writer recognition. This method is based on finding the contours of a handwritten image and extracting a set of chain code based histograms at the global as well as local levels

The main contribution of this paper is that, we have proposed a writer based handwritten document image retrieval using contourlet transform. Unique properties of CT like directionality and anisotropy made it a powerful tool for feature extraction of images in the database. Handwritten document matching was performed using the Euclidean and Canberra distance. Superiority of Canberra distance is observed over Euclidean distance in term of average retrieval rate. The experimental results of proposed method were satisfactory and give better results as compared with earlier approaches. The rest of paper is organized as follows. Section II, discusses the overview of the proposed system. Feature extraction phase is presented in section III. Section IV, discusses handwritten document image retrieval phase. In section V the experimental results are presented and finally section VI concludes the work.

II. Overview of the System

The objective of the proposed work is to study the use of edge and texture orientations as image features in image retrieval. The basic architecture of the system is shown in Fig.1. An improved method based on contourlet transform is proposed in this work. There are two issues in building the proposed system:

Manuscript received August 30, 2012; revised October 10, 2012.

M. S. Shirdhonkar is with the Dept. of Computer Science and Engineering, B. L. D. E. A's College of Engineering and Technology, Bijapur, India.(e-mail:ms_shirdhonkar@rediffmail.com)

Manesh B. Kokare is with the Dept. of Electronics and Telecommunication,S. G. G. S, Institute of Engineering and Technology, Nanded, India.(e-mail:mbkokare@sggs.ac.in)

DOI: 10.7763/IJMO.2012.V2.212

1) Every image in the image data base is to be represented efficiently by extracting significant features.
2) Relevant images are to be retrieved using similarity measures between query and every image in the image database.

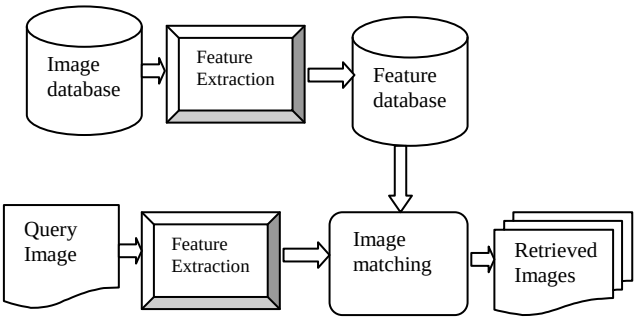


Fig. 1. System architecture for the proposed system

The performance of the proposed system can be tested by retrieving the desired number of the handwritten document images from the document image database. The average retrieval rate is the main performance measures in the proposed system.

III. Feature Extraction Phase

Multiscale and time-frequency localization of an image is offered by wavelets. But, wavelets are not effective in representing the images with smooth contours in different directions. Counterlet Transform (CT) addresses this problem by providing two additional properties viz., directionality and anisotropy [9].which are defined as:
Directionality: The representation should contain basis elements oriented at a variety of directions, much more than the few directions that are offered by wavelets.

Anisotropy: To capture smooth contours in images, the representation should contain basis elements using a variety of elongated shapes with different aspect ratios.

Contourlet transform is a multiscale and directional representation that uses first a wavelet like structure for edge detection and then a local directional transform for contour segment detection.In the double filter bank structure, Laplacian Pyramid (LP) is used to capture the point discontinuities and then followed by a directional filter bank (DFB), which is used to link these point discontinuities into linear structures. The contourlets have elongated supports at various scales, directions and aspect ratios. This allows contourlets to efficiently approximate a smooth contour at multiple resolutions. In the frequency domain, the contourlet transform provides a multiscale and directional decomposition. Contourlet transform is simple and flexible but it introduces redundancy (up to 33%) due to the LP stage. These properties of CT i.e. directionality and anisotropy made it a powerful tool for content based image retrieval.

A. Laplacian Pyramid Decomposition

To obtain multiscale decomposition LP is used. LP decomposition at each level generates a down sampled low pass version of the original image and difference between the original and the prediction that results in a band pass image. The LP decomposition is shown in Fig.2. In LP decomposition process, H and G are one dimensional low pass analysis and synthesis filters respectively. M is the sampling matrix. Here, the band pass image obtained in LP decomposition is then processed by the DFB stage. LP with orthogonal filters provides a tight frame with frame bounds equal to 1. In LP decomposition of an image, $f(i, j)$ represent the original image and its low pass filtered version is $f\,\mathrm{lo}(i, j)$ and the prediction is $f'(i, j)$. The prediction error is given by

$$P\mathrm{e}(i, j) = f(i, j) - f'(i, j) \quad (1)$$

The directional decomposition is performed on $P\mathrm{e}(i, j)$ as it is largely decorrelated and requires less number of bits than $f(i, j)$. In equation (1), $P\mathrm{e}(i, j)$ represents a band pass image. Further decomposition can be carried by applying equation (1) on $f\,\mathrm{lo}(i, j)$ iteratively to get $f\,\mathrm{l}1(i, j)$, $f\,\mathrm{l}2(i, j)\ldots f\,\mathrm{ln}(i, j)$, where 'n' represents the number of pyramidal levels. In LP reconstruction the image is obtained by simply adding back the difference to the prediction from the coarse image.

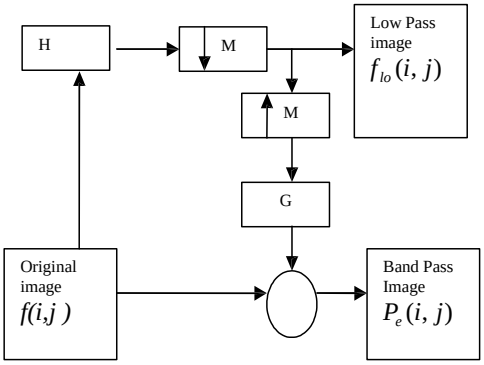


Fig. 2. LP Decomposition (One Level)

B. DFB Decomposition

DFB is designed to capture the high frequency content like smooth contours and directional edges. Several implementations of these DFBs are available in the literature[10].This DFB is implemented by using a k-level binary tree decomposition that leads to 2k directional sub-bands with wedge shaped frequency partitioning as shown in Fig.4. But the DFB used in this work is a simplified DFB , which is constructed from two building blocks. The first one is a two-channel quincunx filter bank with fan filters. It divides a 2-D spectrum into two directions, horizontal and vertical. The second one is a shearing operator, which amounts to the reordering of image pixels. Due to these two operations directional information is preserved. This is the desirable characteristic in system to improve retrieval efficiency. Band pass images from the LP are fed to DFB so that directional information can be captured. The scheme can be iterated on the coarse image. This combination of LP and DFB stages result in a double iterated filter bank structure known as contourlet filter bank, which decomposes the given image into directional sub-bands at multiple scales

IV. Handwritten Document Image Retrieval Phase

A. Feature Database Creation

To conduct the experiments, each image from database is decomposed using CT with a 4 level (0, 2, 3, 4) LP decomposition. At each level, the numbers of directional subbands are 3, 4, 8 and 16 respectively. These parameters results in a 32 dimensional feature vector. To construct the feature vectors of each image in the database, the Energy and Standard Deviation (STD) were computed separately on each subband and the feature vector was formed using these two parameter values and normalized. The retrieval performance with combination of these two feature parameters always outperformed that using these features individually. The Energy and Standard Deviation of subband is computed as follows

$$E_k = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} |W_k(i, j)| \quad (2)$$

$$\sigma_k = \left[\frac{1}{M \times N} \sum_{i=1}^{N} \sum_{j=1}^{M} (W_k(i, j) - \mu_k)^2 \right]^{\frac{1}{2}} \quad (3)$$

where $W_k(i, j)$ is the k^{th} contourlet decomposed subband, $M \times N$ is the size of contourlet decomposed subband, and μ_k is the mean of the k^{th} subband. The resulting feature vector using energy and standard deviation are $\bar{f}_E = [E_1 \quad E_2 \quad ... \quad E_n]$ and $\bar{f}_\sigma = [\sigma_1 \quad \sigma_2 \quad ... \quad \sigma_n]$ respectively. So combined feature vector is

$$\bar{f}_{\sigma\mu} = [\sigma_1 \quad \sigma_2 \quad ... \quad \sigma_n \quad E_1 \quad E_2 \quad ... \quad E_n] \quad (4)$$

We normalized the feature vector and by applying the following statistical normalization method given in Eq. (5) and (6) respectively.

$$\bar{f}_{VE} = \frac{\bar{f}_E - \mu_{\bar{f}_E}}{\sigma_{\bar{f}_E}} \quad (5)$$

$$\bar{f}_{V\sigma} = \frac{\bar{f}_\sigma - \mu_{\bar{f}_\sigma}}{\sigma_{\bar{f}_\sigma}} \quad (6)$$

where $\mu_{\bar{f}_E}$, $\mu_{\bar{f}_\sigma}$ and $\sigma_{\bar{f}_E}$, $\sigma_{\bar{f}_\sigma}$ are the mean and the standard deviation of $\bar{f}_E$, $\bar{f}_\sigma$ respectively. Finally, the resultant feature vector will be the combined normalized vector f_V .

$$f_V = [\bar{f}_{VE} \quad \bar{f}_{V\sigma}] \quad (7)$$

For the creation of feature database, the above procedure is repeated for all the images in the database and these feature vectors are stored in the feature database

B. Document Matching

There are several ways to work out the distance between two points in multidimensional space. The most commonly used is the Euclidean distance measure. It can be considered the shortest distance between two points. We have used Canberra distance and Euclidean distance as similarity measure. If x and y are the feature vectors of the database and query image respectively, and have dimension *d*, then the Canberra distance is given by Eq. 8

$$\text{Canb}(x, y) = \sum_{i=1}^{d} \frac{|x_i - y_i|}{|x_i| + |y_i|} \quad (8)$$

and Euclidean distance is given by Eq. 9.

$$Ed(x, y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2} \quad (9)$$

The average retrieval rate for the query image is measured by counting the number of handwritten document images from the same category which are found in the top 'N' matches. In this work, retrieval performance of the proposed method is compared using Euclidean distance and Canberra distance as similarity measures. Euclidean distances are always not the best similarity measure. Because, the distances in each dimension are squared before summation, results in large dissimilarity. It is observed that results with Canberra distance are superior over Euclidean distance of average retrieval rate.

V. Experimental Results

A. Image Database

The handwritten document images were collected using either black ink or blue ink (No pen brands were taken into consideration), on a white A4 sheet of paper, with one handwriting document image per page. A scanner subsequently digitized the document, contained on each page, with a resolution of 256 x 256 pixels. A number of experiments were carried out to show the effectiveness of the proposed system. A group of 10 writers are selected for 16 specimen handwritten document which make the total of 10x16=160 handwritten document database.

B. Retrieval Performance

For each experiment, one image was selected at random as the query image from each writer and thus retrieved images were obtained. Then the users asked to identify those images that are related to their expectations from the retrieved images. Table1 resumes the precision rates for different image requests of the database.

TABLE I: Average Precision and Recall Values for Database Using Two Distance Measure

	CT using Euclidean distance		CT using Canberra Distance	
Number of Top matches	Precision %	Recall %	Precision %	Recall %
Top 1	100	6.25	100	6.25
Top 2	95	11.87	95	11.87
Top 5	74	23.25	94	25.6
Top 8	66	33.13	76	38.10
Top 10	62	38.75	76	47.50
Top 15	50	46.87	63	58.75
Top 20	43	53.75	57	71.25

The precision is defined as the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. Results correspond to precision and recall rate for a Top1, Top 2, Top 5, Top 8, Top 10, Top 15 and Top 20 and Comparative retrieval performance of the proposed system on the database using CT features is shown in Table 1. Retrieval performance of the proposed method is compared using Euclidean distance and Canberra distance as similarity measures. The Comparative performances in terms of average retrieval rate are shown in Fig. 3. To retrieve images from the database those have a similar writing style to the original request. In Fig.4, retrieval example results are presented in a list of images having a query image

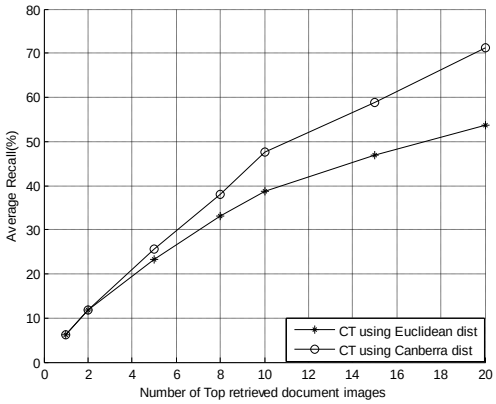
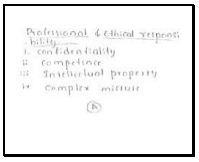


Fig. 3. Comparative average retrieval rate



Query document image

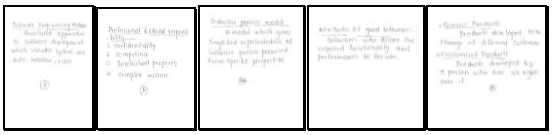


Fig. 4 List of the five most similar retrieved handwritten documents images from the database

VI. Conclusions

We have described a new approach towards writer based handwritten document image retrieval using Contourlet transform. This approach is based on the observations of the handwriting of different writer is visually distinctive and global approach based on texture analysis has been adopted. Features were extracted from handwriting images using Contourlet Transform technique. Handwritten document matching was performed using the Euclidean and Canberra distance. Superiority of Canberra distance is observed over Euclidean distance in term of average retrieval rate. Retrieval results with proposed method are very promising with precisions and recalls.

Acknowledgment

The authors would like to appreciate all participants who gave permission to use their handwritten documents in this study.

References

[1] S. Srihari, S. Shetty, S. Chen, H. Srinivasan, and C. Huang, "Document Image Retrieval using Signatures as Queries," in *Proc. of the Second International Conference on Document ImageAnalysis for Libraries* (DIAL'06),2006.

[2] B. Zhang, S. N. Srihari, "Binary Vector Dissimilarly Measures for Handwriting Identification, In Document Recognition and Retrieval," *SPIE*, pp. 28-38, 2003

[3] L. Schomaker, and M. Bulacu, "Automatic Writer Identification Using Connected Component Contours and Edge Based Features of Uppercase Western Script," *IEEE Trans. of Pattern Analysis and Machine Intelligence* , vol. 26, no. 6, pp. 787-798, 2004.

[4] A. Bensefia, T. Paquet, and L. Heutte, "A Writer Identification and Verification System," *Pattern Recognition Letters*, vol .26, issue 13, pp. 2080-2092, 2005.

[5] R. Pareti and N. Vincent, "Global Method Based on Pattern Occurrences For Writer Identification," in *Proc. of the 10 th International Workshop on Frontiers in Handwriting Recognition La Baule*, France, 2006.

[6] Guillaume Joutel and Veronique Eglin, Stephane Bres, Hubert Emptoz, "Curvelets Based Features Extraction of Handwritten Shapes for Ancient Manuscript Classification," *SPIE*, vol. 6500, pp. 1-10, 2007.

[7] Imran Siddiqi and Nicole Vincent, "Combining Global and Local Features for Writer Identification in Handwritten Documents," in *Proc. of the 11th International Conference on Frontiers in Handwriting Recognition*, Canada,2008.

[8] Imran Siddiqi and Nicole Vincent, "A Set of Chain Code Based Features For Writer Recognition," *10th International Conference on Document Analysis and Recognition*, pp. 981-985, 2009

[9] Ch. Srinivasa Rao, S. Srinivas kumar, and B. N. Chatterji, "Content Based Image Retrieval Using Contourlet Transform," in *ICGST-GVIP Journal*, vol. 7, issue 3, Nov.2007.

[10] Duncan D. Y. Po and Minh N Do. "Directional Multiscale Modeling of Image Using the Countourlet Transform," *IEEE Transactions on Image Processing*, vol. 15. no. 6, pp. 1610-1620, 2006 .

M. S. Shirdhonkar completed his B. E., and M.E. from the Department of Computer Science and Engineering, Shivaji University, Kolhapur, India in the years 1994, 2005 respectively. He is now pursuing his Ph.D from S.R.T.M. University, Nanded, Maharastra, India. His area of research interest includes image processing, pattern recognition, and document image retrieval. He is a life member of Indian Society for Technical Education and Institute of Engineers

Dr. Manesh Kokare received the Diploma in Industrial Electronics Engineering from Board of Technical Examination, Maharashtra, India, in 1990, and B.E. and M. E. Degree in Electronics Engineering from Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIE&T) Nanded, Maharashtra, India, in 1993 and 1999 respectively, and Ph.D. from the Department of ECE, Indian Institute of Technology, Kharagpur, India, in 2005. Since June1993 to Oct1995, he worked with Industry. From Oct 1995, he started his carrier in academics in the Department of Electronics and Telecommunication Engineering at SGGSIE&T, Nanded, where he is presently holding position of Associate Professor. He has published more than 50 papers in international and national journals and conferences. He received the prestigious **Career Award for Young Teachers (CAYT)** for the year 2005 from All India Council for Technical Education (AICTE), New Delhi, India. In December 2009, he honored with "Best Achiever of SGGSIE&T Alumni". Recently Dr. Kokare has been awarded "BOYSCAST" Fellowship for the year 2010-2011 by the Department of Science and Technology, Government of India to carry out his Post Doctoral research work at University of California Santa Barbara, USA. He is a life member of System Society of India, Indian Society for Technical Education, Institution of Electronics and Telecommunication Engineers, and Member of IEEE.