The Effect of SIFT Features as Content Descriptors in the Context of Automatic Writer Identification in Malayalam Language

Sreeraj.M
Department of Computer Science
Cochin University of Science and Technology
Cochin, India
msreeraj@cusat.ac.in

Sumam Mary Idicula
Department of Computer Science
Cochin University of Science and Technology
Cochin, India
sumam@cusat.ac.in

***Abstract*—The span of writer identification extends to broad domes like digital rights administration, forensic expert decision-making systems, and document analysis systems and so on. As the success rate of a writer identification scheme is highly dependent on the features extracted from the documents, the phase of feature extraction and therefore selection is highly significant for writer identification schemes. In this paper, the writer identification in Malayalam language is sought for by utilizing feature extraction technique such as Scale Invariant Features Transform (SIFT).The schemes are tested on a test bed of 280 writers and performance evaluated.**

***Index Terms*— Codebook, Feature extraction, Scale Invariant Features Transform (SIFT), Malayalam, Writer identification.**

I. INTRODUCTION

The scope of writer identification is becoming more prominent these days due to the usage of the same in digital rights administration, forensic expert decision-making systems, and document analysis systems. It is also used as a strong tool for physiological identification purposes. The writer identification is also utilized in authentication system by combining with the writer verification for many fields of confidential data handling purposes. The motivation for Malayalam writer identification scheme stems out from the challenges like (i) Meager Allographic Variation of writers in Malayalam language (ii) Insufficient discriminating capacity of a single character in Malayalam language (iii)Non-existence of uppercase and lowercase in Malayalam language (iv) Absence of dataset [1][2].

Any writer identification scheme generally consists of training phase and identification phase. The identification rate of a writer identification scheme is highly dependent on the feature selection and feature extraction phases. The characteristics of Malayalam script should be taken into account while doing the feature selection. The resultant feature vectors are used to generate the codebook of each writer. Also this vector is used for both training and identification phases. In this paper, feature selection and extraction methodology, namely, Scale Invariant Features Transform (SIFT) are followed.

SIFT features have been extensively utilized in pattern recognition and classification mostly in object recognition. There is a rich track record of usage of SIFT features in robust digital watermarking [3], face authentication [4], graffiti tags recognition [5], car make and model recognition [6] and fingerprint verification [7]. Local features based on SIFT are somewhat invariant to many of the sources of variability. Locality is important, because even if occlusions obscure some portions of the handwriting, there may be enough local features extracted elsewhere which will help to classify the image correctly. This is one of the main motivation to perform the study of SIFT scheme in Malayalam handwritings as they have meager allographic variation. In this situation, SIFT with its immunity to variation could give better results in feature extraction of Malayalam handwritings. This work can be considered as a first attempt with Malayalam handwritings, though studies have been done in the spatial domain [1] [2].

The paper elaborates methods developed to incorporate SIFT scheme to extract writer dependent traits for Malayalam handwritten documents. The results from the tests are further analyzed in the later section.

In this paper, Section II details a design of the writer identifications scheme for Malayalam language. Section III outlines the design of the system developed for implementing the writer identification technique. Section IV describes the feature extraction techniques of SIFT. Section V gives the details of codebook generation. Section VI provides the implementation details. Section VII discusses the test and measurements used for the result analysis. Section VIII analyses the results obtained and provide valid conclusions. The paper is concluded in Section IX.

II. RELATED WORK

A comprehensive review covering the research work in automatic writer identification until 1989 is given in [8]. An extension including work until 1993 has been published in [9]. The on-line handwritten data contains more information about the writing style of a person such as speed, angle, or pressure that is not available in the offline data. Thus, the online classification task is considered to be less difficult than the

978-1-4673-2015-3/12/$31.00 ©2012 IEEE

offline ones [10] [11]. Further, writer identification can also be divided into two parts: text-dependent and text-independent writer identification. This classification is mainly on the different feature level such as character, word, line, paragraph and the document level. Srihari et al. [12] [13] [14] propose a large number of features based on two categories such as Macro level and micro level operate at document/paragraph/word level of CEDAR database using multi-layer perceptron and achieved 98% accuracy. Bensefia et al. [15] [16] [17] [18] used graphemes generated by a handwriting segmentation method to encode the individual characteristics of handwriting independent of the text content. Also Bulacu et al. presented text-independent Arabic writer identification by combining some textural and allographic features [19] [20].

III. System Architecture

In the system architecture (Fig.1), preprocessing, feature extraction and codebook generation are the phases used for training and identification of the writer identification scheme.

A. Preprocessing

The main purpose of this is to remove unwanted areas and noise of the raw input image as it will affect the processing in subsequent phases. The naive image of documents should be pre-processed at first hand through the following steps.

1. Calculate the threshold value by selecting the image mean and the standard deviation and normalize.
2. Convert the documents to grey-scale image using the above threshold.
3. Image de-noising is practiced to attain the perfect binary image of the documents [21].

B. Segmentation

In this module the whole document of writer is divided into different zones to capture their individuality and total no. of zones can be varied from 4 to 6 depending on the aspect ratio of the document.

IV. Feature Extraction

This phase is the essence of the system. Here we incorporate technique and scheme to extract the individuality features attributed to a writer.

A. Scale Invariant Features Transform (SIFT)

SIFT is used to extract distinctive invariant local features from images [22]. Local features are extracted from the preprocessed documents. The local features will include scale, location and orientation of the document image. These features are then utilized to detect keypoints of all sizes. The scale and location feature of a document is defined as the function L(x,y,α),which is convolution of a variable scale Gaussian G(x,y,α) with an input document image I(x,y) as follows [9]:

$$L(x, y, \alpha) = G(x, y, \alpha) * I(x, y) \tag{1}$$

where * is the convolution in the x and y directions

$$G(x, y, \alpha) = \frac{1}{\left(2\pi\alpha^2\right)^{\frac{1}{2}}} \exp\left(-\frac{x^2 + y^2}{2\alpha^2}\right) \tag{2}$$

The difference between two nearby scales is given by Difference of Gaussian function, where certain locations are discarded from the candidate keypoints, when compared with nearby data. It is also to be noted that a location possess more than one orientation. To determine the dominant direction, a gradient histogram is computed with respect to the neighbourhood. Peaks at the histogram are correspondent with dominant orientation. The dominant orientation of each keypoint is obtained by computing the gradient magnitude M(x,y) and orientation θ(x, y)of the scale space for the scale of that keypoint:

$$M(x, y) = \sqrt{(K(x+1, y) - K(x-1, y))^2 - (K(x, y+1) - K(x, y-1))^2} \tag{3}$$

$$\theta(x, y) = \arctan \frac{K(x, y+1) - K(x, y-1)}{K(x+1, y) - K(x-1, y)} \tag{4}$$

Local image gradients are measured at the selected scale in the region around each key point and transformed into a representation that allows local shape distortion and change in illumination. An array of 4*4 smoothed histograms in neighbourhood pixels of the keypoint selected is quantized to 8 orientation bins, to represent important aspects of the region. Thus it gives 16*8 =128-dimension SIFT vector which has then to be normalized.

V. Code book generation

This phase is the essence of the system. Here we incorporate technique and scheme to extract the individuality features attributed to a writer.

The codebook- redundant writing patterns- could generally be writer specific, where the frequent writing forms are extracted separately for each writer. While generating a codebook of the characteristic writings, we group them into classes which are represented by a set of features. Methods like k-means, fuzzy c-means, learning vector quantization and the closely related self organizing maps have been successfully applied to similar problems of clustering allograph or graphemes [23] [24].

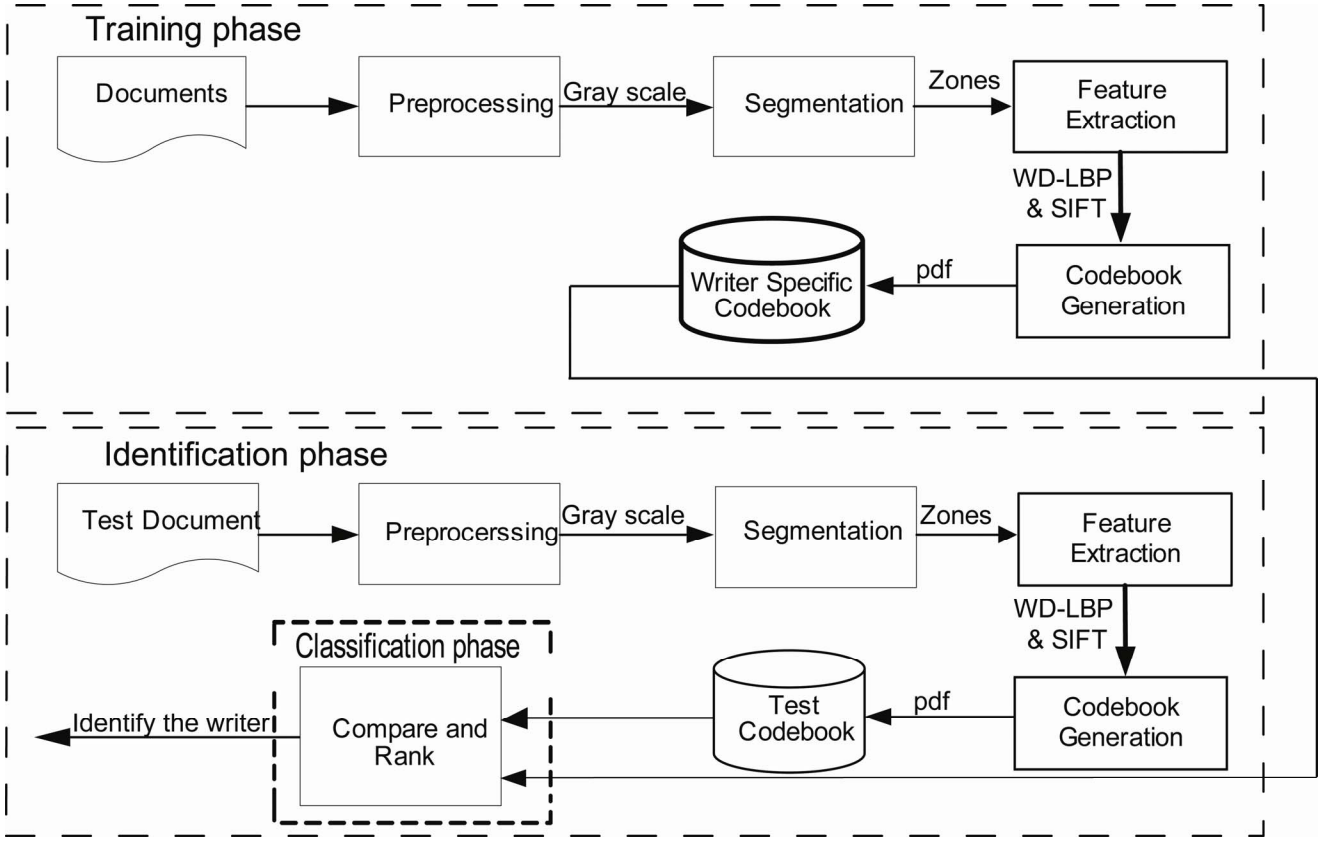

Fig. 1 System Architecture

(**Generating the writer specific code book**):

1. Split each document into equal sized zones. Total no. of zones is set to be 6 for SIFT feature.
2. Extract key point descriptors from each zone.
3. Calculate the probability density function (pdf) of each descriptor over the zones to obtain a fused feature vector, which is a representative of a zone.
4. Cluster the input documents based on the above fused feature vector using Kohonen self organization map (SOM 2D).
5. Obtain the writer specific feature vector by averaging the representative feature vector of all the zones and store it as a writer specific codebook entry.

VI. Implementation

The handwriting samples of 280 different users of similar as well as variable content of Malayalam text were collected. The images have been scanned at 400 dpi, 8 bits/pixels, gray-scale. A total of 280 writers contributed to the data set with 50 writers having only one page, 215 writers with at least 2 and 15 writers with at least 4 pages. Each page consists of 21 lines of words with a minimum 30 characters in each line. We kept only the first two images for the writers having more than two pages and divided the image into two parts for writers who contributed a single page thus ensuring two images per writer, one used in training while the other in testing.

All these images were preprocessed as mentioned in Section 2 before doing the feature extraction. The feature extraction is done by using SIFT. The feature vector is then used for training as well as in identification phase. In the identification phase, steps adopted as follows.

1. Perform the nearest-neighbour classification in a "leave-one-out" strategy between the writer descriptor of test document (query descriptor) and writer specific codebook of trained documents and assign a label c of writer to the query descriptor using

$$c = \arg\min_{i} \left(\chi^2\left(C^i, Q\right)\right), i \in \{1,2,.....,N\} \quad (5)$$

Both χ square distance and Euclidean distances are experimented to measure the variability between them. The image distance between document **A** and document **B** is given by Euclidean distance from the i[th] keypoint in document **A** to all the keypoints of document **B**.

$$D(A,B) = \frac{1}{\mathrm{K}_a} \sum_{i=1}^{\mathrm{K}_a} D\left(A_i, B\right) \quad (6)$$

where *K*a is the number of keypoints in document A.

2. Prepare a sorted hit list with increasing distance value between the query descriptor and writer specific codebook.
3. Select the first ranked sample which will ideally identify the writer

VII. Tests and Measurements

The results of the implementation are measured on the basis of efficiency of features and classifiers, stability test on features and classifiers and overall performance.

A. Efficiency

Efficiency is measured using identification rate where, identification rate of this system is calculated by choosing the successfully identified writers without any false positive.

B. Stability test

This test is conducted to check the stability of the features and classifiers in writer identification. This is done by analysing the performance of features at different reference points /zones (one word – two words; one line - two lines – three lines – four lines; one paragraph; full page.) on the documents of the total 280 writers.

C. Overall performance

When the writer identification is considered as a system, the commonly used classification functions in data mining are sensitivity, specificity, precision and F-measure.

VIII. Result Analysis

The aim of the system is to identify a decisive feature for writer identification for Malayalam documents. SIFT features based on efficiency, stability and consistency were measured and these are compared to achieve the aim. The comparison further extended to classifiers, Naïve-Bayes, k-NN, SVM and Adaboost M2 to identify the most ideal classifier for Malayalam characters.

A. Efficiency

The feature extraction is done using the SIFT, and its efficiency in identifying writers is measured with respect to the classifiers, Naïve-Bayes, k-NN, SVM and Adaboost M2 and the observations are given in Table I.

As shown in the Table I, as the number of writers increases, the performance of Naive-Bayes is deteriorating faster than the other classifiers.

In computing SIFT features, the variability between sample and codebook distributions is computed in terms of chi-square distance and Euclidean distance. It is obvious from the Fig. 2 that both of them perform same for less number of writers. But as the number of writers' increases, better results are obtained when we depend on chi-square distance.

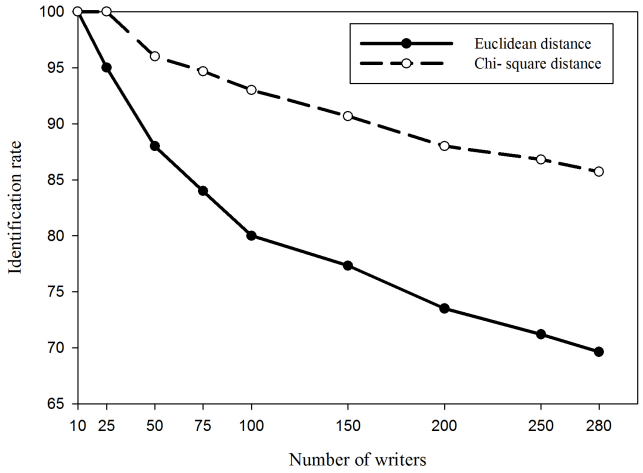

Fig. 2 Comparative results of different distances used for SIFT features

TABLE I
PERFORMANCE OF SIFT FEATURES WITH DIFFERENT CLASSIFIERS

Number of writers	Naïve Bayes	k-NN	SVM	Adaboost M2
10	100	100	100	100
25	96	96	96	100
50	92	94	94	96
75	90.67	92	93.33	94.67
100	88	90	92	93
150	84.67	89.33	90	90.67
200	82	86.5	87	88
250	79.6	84.8	85.2	86.8
280	78.21	83.57	84.29	85.714

B. Stability test

Fig.3 depicts the stability of SIFT feature in each classifiers namely Naïve bayes, k-NN, SVM and Adaboost M2. It is seen that the zone with maximum amount of text as one complete page yielded greater identification rates as compared with zones having less text in the different classifiers. Also all classifiers except Naïve bayes produced nearly similar results.

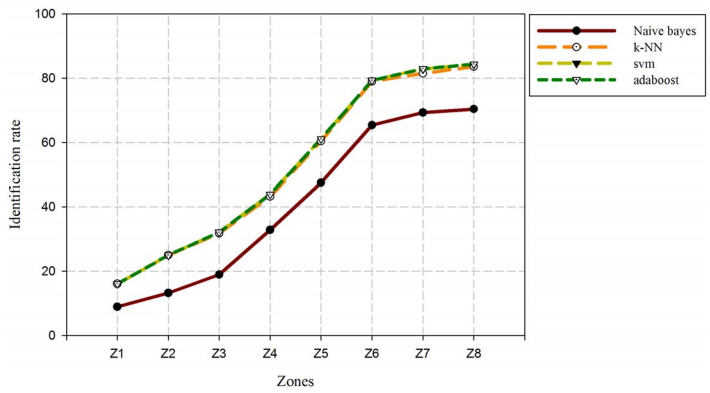

Fig. 3 Performance with respect to amount of text

C. *Overall performance*

The overall performance of the system developed was measured using sensitivity, specificity, precision and F-measure functions and is depicted in Table II.

TABLE II
OVERALL PERFORMANCE

	sensitivity	specificity	precision	F measure
Naïve bayes	0.7300879	0.9991645	0.7803571	0.7386343
k-NN	0.7954252	0.9993765	0.84375	0.7986281
SVM	0.8019841	0.9993965	0.8428571	0.8082375
Adaboost M2	0.8224887	0.999457	0.8625	0.8271695

IX. CONCLUSION

In this paper, the writer identification scheme in Malayalam language is sought for by seeking the feature of SIFT. SIFT is used to extract distinctive invariant local features like scale, space and orientation from images. The entire scheme is tested on a total of 280 writers, wherein 50 writers with one page, 215 writers with at least 2 and 15 writers with at least 4 pages. An identification rate of 85.714% is obtained by using SIFT feature. In our experiments SIFT shows more efficient, consistent and stable performance.

REFERENCES

[1] M. Sreeraj, and Sumam Mary Idicula, "Identifying Decisive Features for Distinctive Analysis of Writings in Malayalam", International Magazine on Advances in Computer Science and Telecommunications (IMACST), Association for Computer Science and Telecommunications (AKOMNAT TEL DOO), Vol.2 , No.1, pp. 13-20, May,2011.

[2] Sreeraj, M., Sumam Mary Idicula, "A Novel Approach to Writer Identification in Malayalam using Graphemes", in Proc.5th International Conference on Information Processing,ICIP 2011, Bangalore, India, Communications in Computer and Information Science(CCIS), Springer-Verlag, Vol. 157, pp.646-651, Aug 5-7,2011.

[3] H. Kim, H. Lee, and H. K Lee, "Robust Image Watermarking using Local Invariant Features", Proceedings in Proc. SPIE, vol. 45, no. 3, 2006.

[4] G. Enrico, B. Manuele, L. Anderea and T. Massimo, "On the use of SIFT features for face authentication", in Proc. Conference on Computer Vision and Pattern Recognition Workshop, pp. 91–110, 2006.

[5] P. Schwarz, "Recognition of Graffiti", BS Thesis, The University of Western Australia, 2006.

[6] L. Dlagnekov, "Video-based Car Surveillance: Licence plate, Make and Model Recognition," MSc Thesis, University of California, San Diego, 2005.

[7] U. Park, S. Pankanti and A.K. Jain, "Fingerprint Verification using SIFT Features". In: Proc. of SPIE Defense and Security Symposium, Orlando, Florida (2008).

[8] R. Plamondon and G. Lorette, "Automatic Signature Verification and Writer Identification—The State of the Art," Pattern Recognition, vol. 22, no. 2, 1989, pp. 107-131.

[9] F. Leclerc and R. Plamondon, "Automatic signature verification: The state of the art—1989–1993," Int. J. Patt. Recognit. And Artificial Intell., vol. 8, no. 3, June 1994, pp. 643–660,.

[10] A. Schlapbach, L. Marcus, H. Bunke, "A writer identification system for on-line whiteboard data", Pattern Recognition Journal 41 (2008) 23821–23897.

[11] L. Schomaker, "Advances in Writer identification and verification", in: Ninth International Conference on Document Analysis and Recognition (ICDAR), 2007.

[12] S. Srihari, S. Cha, H. Arora, and S. Lee, "Individuality of Handwriting," J. Forensic Sciences, vol. 47, no. 4, July 2002, pp. 1-17.

[13] S. Srihari, M. Beal, K. Bandi, V. Shah, and P. Krishnamurthy, "A Statistical Model for Writer Verification," Proc. Eighth Int'l Conf. Document Analysis and Recognition (ICDAR), 2005,pp. 1105-1109.

[14] S. N. Srihari, S.-H. Cha, and S. Lee, "Establishing Handwriting Individuality Using Pattern Recognition Techniques," in Proceedings of the Sixth International Conference on Document Analysis and Recognition, 2001, pp. 1195-1204.

[15] A. Bensefia, T. Paquet, and L. Heutte, "A Writer Identification and Verification System," Pattern Recognition Letters, vol. 26, no. 10, Oct. 2005, pp. 2080-2092.

[16] A. Bensefia, T. Paquet, and L. Heutte, "Handwritten Document Analysis for Automatic Writer Recognition," Electronic Letters on Computer Vision and Image Analysis, vol. 5, no. 2, Aug. 2005, pp. 72-86.

[17] A. Bensefia, A. Nosary, T. Paquet, and L. Heutte, "Writer Identification by Writer's Invariants," Proc. Eighth Int'l Workshop Frontiers in Handwriting Recognition, Aug. 2002, pp. 274-279.

[18] A. Bensefia, T. Paquet, and L. Heutte, "Information Retrieval Based Writer Identification," Proc. Seventh Int'l Conf. Document Analysis and Recognition (ICDAR), Aug. 2003, pp. 946-950.

[19] M. Bulacu, L. Schomaker,A.Brink, "Text-independent writer identification and verification on offline Arabic handwriting, in:Ninth Conference on Document Analysis and Recognition(ICDAR),2007.

[20] M. Bulacu,L. Schomaker, "Text-independent writer identification and verification using textural and allographic features", IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)29(4)(2007)701–717 Special Issue—Biometrics: Progress and Directions.

[21] A. Al-Dmour, R. AZitar, "Arabic writer identification based on hybrid spectral statistical measures", J. Experimental and Theoretical Artificial Intelligence, vol 19(4), pp. 307-332 (2007).

[22] D. Lowe, "Distinctive Image features from Scale- invariant Keypoint"., J. Computer Vision., vol. 60, no. 2, pp. 91–110, 2004.

[23] G. X. Tan, C. Viard-Gaudin and A. C. Kot. "Automatic writer identification framework for online hand written documents using character prototypes", J. Pattern Recognition, 42,pp. 3313-3323, 2009

[24] F. Chang, C. H. Chou, C. C. Lin and C. J. Chen, "A prototype classification method and its application to handwritten character recognition", In: Proc. IEEE International Conference on Systems, Man and Cybernetics, 2004.