Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 5, May 2015, pg.1057 – 1061

RESEARCH ARTICLE

Writer Recognizer for Offline Text Based on SIFT

Priyanka Kathe
Department of Computer Engineering,
MET BKC IOE,
Nashik, India
Email: kathepriyanka3@gmail.com

Vaibhav Dabhade
Department of Computer Engineering,
MET BKC IOE,
Nashik, India
Email: vaibhavdabhade@rocketmail.com

Abstract— The writer recognizer for offline text method is based on scale invariant feature transform (SIFT) descriptor, composed of training, enrollment, and identification stages [1]. The handwriting images are segmented to word regions (WRs) in all the stages using an filter. Then, the SIFT descriptors (SDs) of WRs and the corresponding scales and orientations (SOs) are extracted. Training stage composes the clustering of SDs and training samples for constructing SD codebook. In enrollment stage, the SDs of the input handwriting are adopted to form an SD signature (SDS) by looking up the SD codebook and the SOs are utilized to generate a scale and orientation histogram (SOH). In the identification stage, the SOH and SDS of the input handwriting are extracted and matched with the enrolled ones for identification

Keywords— SIFT, Offline text-independent writer recognizer, word segmentation, SIFT descriptor signature, Scale and Orientation histogram.

I. INTRODUCTION

The writer recognizer for offline text is very important for documents authorization, forensic analysis, and calligraphic relics identification. Automated writer recognizer for offline text is to determine the writer of a text among a number of known writers using their handwriting images. Number of extensive researches have been conducted in this field [1]. The structure-based approach as well as texture-based approach are used for writer identification. In the texture-based approaches it takes handwriting texts as a special texture image and extract the textural features for writer identification. Zhu et al [2], Said et al [3], and Hanusiak et al [4] used a grey-level co-occurrence matrix (GLCM) to extract the textual features from the handwriting images. Hanusiak et al [5] also extracted features based on hidden Markov tree. Comparing the textural features with the structural features of handwriting are much more intuitionistic, notable and stable for writer identification. Hence , recently there are large number of the researches are focused on the structure-based approaches for writer identification. In the structure based approaches features are extracted from the points on contours of handwritings. Bulacu et al [5] proposed several edge-based directional features of handwriting, i.e. edge direction distribution, edge hinge distribution and the directional co-occurrence PDF to characterize the individuality of the writer.

II. EXISTING SYSTEM

The existing systems are mostly based on structure-based approaches which contours the allograph fragments of hand-writing, and are easily affected by the slant and aspect ratio of the characters in handwriting [6]. These approaches extract the features from allographs, which fail to extract the structural features between the allographs in the same words. The words are always taken as a whole and the structures of the whole word have a strong discriminability for different writers and are stable when writing a document. For characterizing writers individuality the structures between allographs in the same word are also important. Now to deal with these problems, a scale invariant feature transform (SIFT) descriptor method extracts the key point based structural features at word level from handwriting images, which contains the structural in-formation of whole words and is insensitive to the aspect ratio and slant of the characters .So the automated writer identification for offline text using Scale invariant feature transform (SIFT) descriptor will overcome the failure of the existing system. This SIFT descriptor is used to characterize writers individuality for structures between allographs in the same word.

III. **PROPOSED SYSTEM**

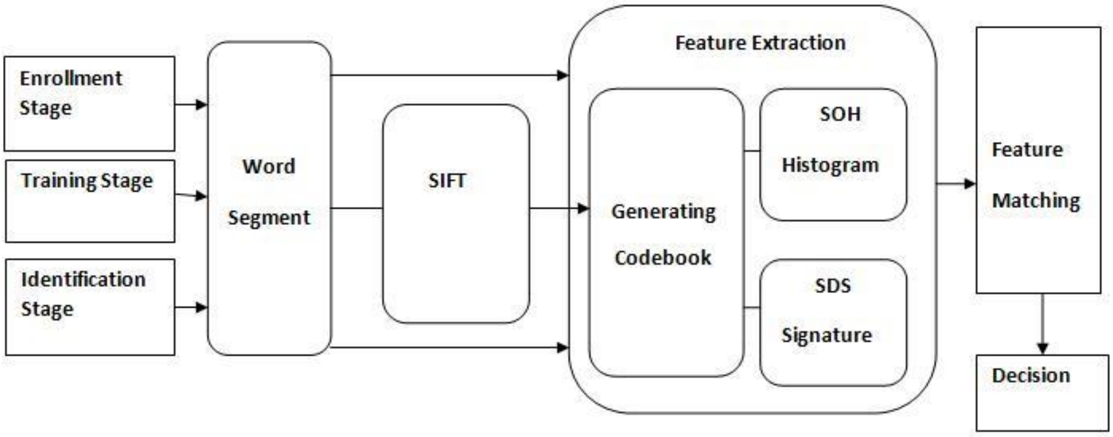


Fig. 1. System Architecture of Proposed System

The proposed system have the training, enrollment, and identification stage as shown in Fig 1. In all the three stages the handwriting image is segmented into word regions (WRs). SIFT is used to detect the key points and extract their SIFT descriptors (SDs), and the corresponding scales and orientations (SOs)from the WRs. The SOs and SDs will be used in different ways in different stages. The SDs extracted from the training data set are used to generate a codebook for the use of identification and enrollment in the training stage. The SO histogram (SOH) and SD signature (SDS), are extracted from SOs and SDs of WRs of the enrolling handwriting image and stored for identification in the enrollment stage. The SOH and SDS are extracted from the input handwriting images and respectively matched with the enrolled ones to get two matching distances, which are then fused to form the final matching distance for decision. As shown in Fig. 1, there are four main parts in the framework ,i.e. word segmentation, codebook generation, feature extraction, and feature matching and fusion.

The writer recognizer for offline text is based SIFT algorithm in order to identify writer individuality among the writers. As shown in the following Fig1. it consist of training,, enrolment, and identification stages. First the handwriting image is segmented into word region using word segment. The generation of codebook is done it uses vectors containing x and y coordinates of the normalized contours can be used to train a clustering algorithm. After training, a specific number of common fragments appearing in peoples handwritings are determined. The results show approximately the same performance for these clustering methods. After construction of the codebook, the feature vector is calculated by an occurrence histogram, each of which corresponds to one codebook member. To construct this histogram, all fragments of the handwriting are first extracted and normalized. Then, for each fragment, the most similar member of the codebook is selected using a Euclidean distance function, and the corresponding is increased by one. Finally, the members of histogram are divided by the sum of them. Using the new method, the number of extracted fragments can be modified with changing the gap parameter. By decreasing this parameter, the number of extracted fragments increases. This increase is especially useful when a short text is available. SIFT are used to detect key points and extract their SIFT descriptor (SDs) as shown in following Fig2. Then feature matching is done in order to get the final result.

A. Word Segmentation

Recent automatic word segmentation techniques are based on text line segmentation. Firstly segmented text lines by using Hough transform, and then segment words from each text line according to the distances between the adjacent connected components in vertical project domain [8]. The word segment process first takes the original handwriting image (I) then convert that original image in binary image (Ib). Filter the binary image into filtered image (If). Then obtain a binary filtered image (Ifb). Semi-word regions (SWRs) are formed. Merge the image (WRs) i.e. form word regions and split overlapping of words.

Generate Codebook

Many word regions (WRs) are obtained from handwriting document image, after word segmentation. To obtain each WR, the SIFT algorithm detects a number of key points and extract their descriptors, scales, and orientations. Fig 2. shows an example of the key points detected in a word region by using SIFT. Large and varying amount of key points from different handwriting images are obtained. The generation of codebook is done it uses vectors containing x and y coordinates of the normalized contours can be used to train a clustering algorithm [10][11] . After training, a specific number of common fragments appearing in peoples handwritings are determined. The results show approximately the same performance for clustering methods

B. Feature Extraction

Since the text in the identifying handwriting document may be totally different with the text in the enrolled handwriting document in offline text-independent writer recognizer system, the layout of the key points may be totally different in the different handwriting images, even if they are written by the same person. Therefore, it will not consider the positions of the key points in the following feature extraction and matching it takes the frequency of each SD and SO occurrences in a handwriting image.

Here the feature of the handwriting images are extracted as follows:

1] SIFT Descriptor Signature (SDS) Extraction :
1. N is the size for SDS feature vector
2. Compute the Euclidean distance as follows

$$ED_{ij} = \sqrt{\sum_{K=1}^{L}(d_{ij} - c_{jk})^2} \quad (1)$$

3. Compute the SDS vector.

(2) Scale and Orientation Histogram (SOH) Extraction :
1. Initialize SOH feature vector with size M.
2. Compute the index in the SOH feature vector
3. Compute the final SOH vector as follows

$$SOH_i = \frac{SOH_i}{\sum_{i=1}^{M} SOH_j}$$

C. Feature Matching Module-

1] The features of two handwriting images are used to calculate their dissimilarity using Manhattan distance as follows

$$D_1(u,v) = \sum_{i=1}^{N} |u_i - v_i| \quad (2)$$

2] The Chi-square distance is calculated which improves the importance of small value components as follows

$$D_2(x,y) = \sum_{j=1}^{M} \frac{(x_j - y_j)^2}{(x_j + y_j)} \quad (3)$$

Then after normalized both $D1$ and $D2$ these two distances are then fused to form a new distance to measure the dissimilarity between two handwriting images as below:

$$D(I1, I2) = w \times D1(u, v) + (1 - w) \times D2(x, y) \quad (4)$$

D. Methodology

The proposed method can be implemented as follows :
1. Word segmentation: The handwriting image is segmented into word regions.

2. Scale invariant feature transform algorithm: It has four major stages of computation:
(a) scale-space construction
(b) Key point localization
(c) Orientation assignment
(d) Key point descriptor extraction.

3. The scale invariant feature transform is used to get the key points of handwriting, their SIFT descriptors, and the corresponding scales and orientations.

4. Codebook generation: Cluster the SDs of the key points extracted from the training samples into N categories and represent each category with its center, which is called a code.

5. SIFT Descriptor Signature extraction and scale and Orientation Histogram extraction is done.

6. Identification stage: Here the feature matching of two handwriting images is done and hence recognition of writers individuality.

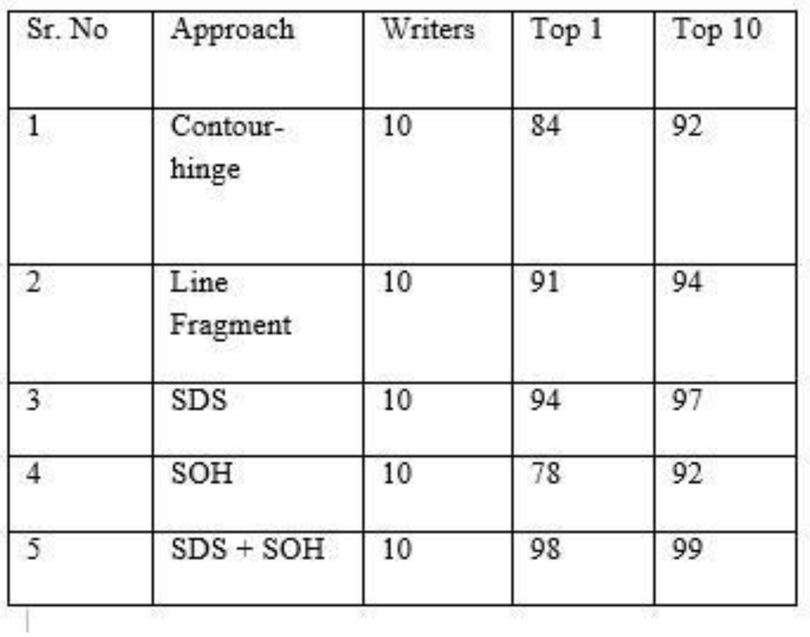
IV. **EXPERIMENTAL SETUP**

The experiment is carried out in java. The handwriting images of different writers are taken as an input. The output is identified writer of the handwriting

V. **RESULTS**

The following Table shows the result for the different methods [1] as follows

TABLE 1. COMPARISON OF METHODS



Sr. No	Approach	Writers	Top 1	Top 10
1	Contour-hinge	10	84	92
2	Line Fragment	10	91	94
3	SDS	10	94	97
4	SOH	10	78	92
5	SDS + SOH	10	98	99

VI. CONCLUSIONS

The goal is to automate the process of writer identification using scanned images of handwriting and thereby to provide a computer analysis of handwriting individuality. In this endeavor, computational factor i.e. SIFT takes center stage the design and use of appropriate representations, computable features capturing the writing style of a person from the scanned handwritten samples. The power of such a representation or feature relies in its ability to maximize the separation between different writers, while remaining stable over samples produced by the same writer. The novel is very effective for automatic writer identification on the basis of scanned images of handwriting. The similarity in handwriting style between any two samples is computed by using appropriate distance measures between their corresponding features. It can also be used for historical document analysis, handwriting recognition system enhancement hand held and mobile devices. To a certain extent its recent development and performance consider as a strong physiologic modalities of identification, such as DNA and fingerprints.

ACKNOWLEDGEMENT

The author is thankful to METs Institute of Engineering Bhujbal Knowledge City Nasik, HOD of computer department, guide and parents for their blessing, support and motivation behind this work.

REFERENCES

[1] H. Said, T. Tan, and K. Baker, “Personal identification based on handwriting,” *Pattern Recognit.*, vol. 33, no. 1, pp. 149–160, Jan. 2000..

[2] L. Schomaker and M. Bulacu, “Automatic writer *identification using connected-component contours and edge-based features of uppercase Western script,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 6,pp. 787–798, Jun. 2004.*

[3] V. Pervouchine and G. Leedham, “Extraction and analysis of forensic document examiner features used for writer identification,” *Pattern Recognit.*, vol. 40, no. 3, pp. 1004–1013, Mar. 2007

[4] M. Bulacu and L. Schomaker, “Text-independent writer identification and verification using textural and allographic features,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 701–717, Apr. 2007

[5] Y. Zhu, T. Tan, and Y. Wang, “Biometric personal identification based on handwriting,” in *Proc. Int. Conf. Pattern Recognit.*, Barcelona, Spain,2000, pp. 797–800.

[6] R. Hanusiak, L. Oliveira, E. Justino, and R. Sabourin, “Writer verification using texture-based features,” *Int. J. Document Anal. Recognit.*,vol. 15, no. 3, pp. 213–226, Sep. 2012.

[7] Youbao Tang and Wei Bu, ”Offline Text independent writer identification based on scale in invariant feature transform” *IEEE Transaction on Information Forensics Security ,VOL. 9,No. 3,March 2014.*

[8] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, “Text line and word segmentation of handwritten documents,” *Pattern Recognit.*, vol. 42, no. 12, pp. 3169–3183, Dec. 2009