Textual data identifications of different writer using Scale Invariant Feature Transform

Abstract
This paper mainly focus on the writer handwritten text identification based on the scale invariant feature transform (SIFT). In this SIFT method three stages followed 1) Training stage 2) Enrolment stage 3) Identification stage. In each stage one default filter is used i.e. Log. This filter used to segment the handwriting text into word regions (WRs). Then the SIFT descriptors (SD's of writing regions and the corresponding scales and orientations (SO) are extracted. In training stage SD code book obtained by clustering the SDs of training samples. In enrolment stage SD signature is used to get SD's of input handwriting by looking up the SD code book and SD are optimised generated a scale and orientation histogram (SOH). In identification stage the SDs & SOH of input handwriting text are extracted and matched with the enrolled ones for identification. At last result on six public data sets (including three English data sets).

1. Introduction
The text identification of writer is to determine the writer text among a number of known writers using handwriting images. The approach existing for offline text independent writer identification can be roughly divided into two categories. 1) Texture based approach. 2) Structure based approach. The texture approach take hand writing texts as a special texture features for writer identification used a grey level co-occurrence matrix (GLCM) extract textual features from the handwriting images. This extracted features based on hidden markow tree (HMT) model wavelet domain for writer identification. The extracted wavelet based textual feature from handwriting images. Then it used Gabor and XGabor filter to extract features from handwriting data and
employed a feature relation graph (FRG) to represent the extracted feature. It considered both local binary pattern (LBP) and local phase quantization (LPQ) as texture descriptors of handwriting for writer verification and identification.

2. Methodology
The new proposed system having 3 stages i.e. Training, enrolment and identification.

Figure 2.1: The Framework of the proposed method
In all these three stages handwriting images segmented into word regions (WRs) Then SIFT used to identify key point and get SIFT descriptors (SDS) & corresponding scales and orientation (SOS) from WRs. The SDS & SOS will be used in different ways in different stage. In the training stage SD’s extracted from the training data set are used to generate code book for the use of enrolment and identification. In enrolment two feature called SD signature (SDs) & SO histogram (SOH) are used from SDs & SOs of WRs of the enrolling handwriting image & stored for identification. In identification stage SDs & SOH are extracted from the input handwriting image and respectively matched with the enrolled once to get two matching distance which are then fused to from final matching distance from decision.

3. Word Segmentation

The handwriting image I converting into word segmentation process can be simply following procedure.
1) Converting I to binary image using OTSU’s algo.
2) Getting all connected components (CS) in I and then computing their average height $h_a$
3) Filtering I, with an isotropic LoG filter to get the filtered image $I_f$
The average height $h_a$ of all Ccs in $I_b$ to decide the variance of the filter as $\sigma = 2.5 * h_a$
4) Binarizing $I_f$ to get a binary image $I_{fb}$ by using threshold obtained by otsu’s algorithm.
5) Assigning each connected component in $I_b$ to the nearest connected regions of $I_{fb}$ to form semi word regions (SWR) which colored different.
6) Merging the SWR’s to get the word regions according to the distance between the adjacent SWR’s.
7) Splitting the overlapping Connected Components runs along multiple text lines from middle line of these boundary box.

SIFT Algorithm has 4 major stages of computation.
1) Scale space construction.
2) Key point localization
3) Orientation assignment
4) Key point descriptor extraction.
In the first stage the original images area decomposed into a Gaussian pyramid and each level of the pyramid called an octave which further decomposed into several sub levels by convolving the initial image at the corresponding pyramid level with DoG filters with different variances. In second and third stage many stable key point are detected and the location, scales and orientations of these key point are computed. In last step SIFT descriptor for each key point is generated.


For each word regions we used SIFT algorithm for detect no of key point and extract their descriptors scales and orientations. In code generation hierarchical kohonen SOm clustering algorithm is used. All of N codes form a SDS code book with size N and based on the code book we will compute a histogram with limited and fixed dimension as feature vector for writer identification.
5. Future Extraction
Sometime the text in the identification hand writing document may be totally different with the text in the enrolled hand writing document in project. At that time the layout of the key point may be totally different in the different hand writing images even if same written by same person. So for mathematical calculation not considering key point considering future extraction and matching by recalculating frequency each SD & SO occurrence in the handwriting image.

6. Conclusion
As according proposed methodology we can identify various writer handwriting using SIFT algorithm which produce various feature accurately and perfectly. The proposed system computes the frequency of local structure feature occurrence in a handwriting image and the local structure of some special stroke make very little contribution to feature extraction so we can expect more and more accuracy in writer identification using SIFT.

References