An Empirical Intrinsic mode based characterization of Indian Scripts

Kavita Bhardwaj*
Indian Institute of Technology
New Delhi,India
kavitab788@gmail.com

Santanu Chaudhury
Indian Institute of Technology
New Delhi,India
schaudhury@gmail.com

Sumantra Dutta Roy
Indian Institute of Technology
New Delhi,India
sumantra.dutta.roy@gmail.com

ABSTRACT

In this paper, we describe a novel technique for Document script identification(DSI) from printed documents, using Empirical Mode Decomposition (EMD). The intrinsic decomposition nature can adaptively decompose script images into a series of modes representing different local features of script images. In this method, Radon transformed script images are decomposed into finite set of IMFs (Intrinsic Mode Functions). The energy concentration in a particular orientation characterises a script texture as it indicates the dominance of individual script in that direction. We demonstrate how the proposed method use these IMFs as feature vectors to distinguish various scripts.

Keywords:

Empirical mode decomposition(EMD), Radon transform, Intrinsic mode function, AdaBoostM1

1. INTRODUCTION

The Identification of the script used in printed documents is useful for the digitization of the conventional paper documents, sorting of document images according to the scripts in which they are written, for selecting appropriate script-specific OCRs for the retrieval of online archives of document images or for indexing of documents in digital library.

Ghosh et al. [1] proposes the categorisation of script recognition methods as structure-based and visual appearance-based. He discussed the methods of both categories at page-level, paragraph-level, word-level and character-level. A vast survey is presented for each of the categories. By reffering Wang et al. [9] and Ghosh et al. [1], it is found that according to the feature extraction, all the methods lying under any category are grouped into three major categories- Statistical-information based methods, Structure-based methods, Texture-based methods. Statistical information-based algorithms use character density distribution, vertical and horizontal projections, for classifying printed documents. Waked et al. [8] used bounding box size distribution, character density distribution, vertical and horizontal projections for the classification of printed documents. Lam et al. [5] has also used statistical features for script identification in printed-documents. These methods are more useful for scripts that differ significantly in style. Structure-based methods focus on extraction and analysis of connected components and use the identification results of these âĂIJsignaturesâĂİ to determine the script(s) used. These methods in general have advantage of discriminating similar scripts. Hochberg et al. [2] exploited the shape characteristics of "textual symbols" for the identification of script(s). Pal and Chaudhuri [6], presented the script characteristics and shape based features for script identification. Visual-appearance and texture analysis-based methods are related, because according to appearance of any text block, corresponding texture analysis-based method can be used for extraction of features. Joshi et al. [4] proposes the Gabor function-based texture analysis to extract features and used hierarchical classification to distinguish among the script(s). Tan [7] developed Gabor function-based texture analysis for machine-printed script identification that discriminates Chinese, Latin, Greek, Russian, Persian, and Malayalam script documents.

For our problem, we propose an algorithm based on Empirical Mode Decomposition(EMD) for textural analysis of script classes. The directionality and periodicity reflect the effective directions for textural processing of subpatterns. Each script class will always exhibit a specific periodicity at a particular angular orientation. This is observed in Radon transformed image of the script classes considered for the problem and they are decomposed in different mode functions to compute directional energy specified by each IMF. The scripts involved in this paper are Devnagari, Roman(English), Malayalam, Bangla and Gurumukhi. The cosine similarity measure is used as our measure to define the most similar script class for the descrimination. We use Adaboost binary decision tree to improve the classification.

The rest of the paper is organized as follows. In (Sec. 2), we summarize the proposed approach and the framework for the problem is described in (subsec. 2.1). The results are shown through Table 3in (Sec. 3). Experimental observation are described in (Sec. 4). Conclusions and future work are discussed in (Sec. 5).

2. THE PROPOSED APPROACH

The method described in this paper involves four main steps.

- First, the preprocessing is performed initially to remove noise that includes binarization of document images.
- Second, Radon transform is computed on document images of each script at different angles of orientation between 0° to 90°. The unique characteristic of each script is observed at a particular orientation in radon transformed image. The

*Corresponding author

DAR '12, December 16, 2012, Mumbai, IN, India

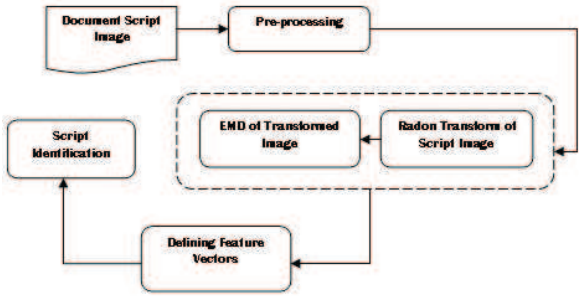

Figure 1: Schematic Block Diagram of the System

transformed image is decomposed using empirical mode decomposition(EMD) in a finite set of IMFs.

- Third, the energy is computed corresponding to each Intrinsic Mode Function(IMF) at different orientations. The energy at a particular orientation characterises a script class texture as it indicates the dominance of individual script class in that orientation.
- This angle of orientation and the IMF exhibiting maximum energy helps to chose a distinguishing feature vector for individual script.
- Fourth, the feature vectors for test script images are obtained similarly and then classified using AdaBoost binary decision tree.

Figure 1 shows a schematic diagram of the system. We will discuss each of the modules in detail in the upcoming sections of this paper.

2.1 A Generalised Framework to Script Identification

We propose a general framework to address the problem of script identification. This proposed scheme (for document script identifcation exploits the textural characteristic of each script. The constructing patterns of each script are formed by oriented linear/curvilinear subpatterns. The energy distributed at different orientations, characterises each script according to the textural characteristics of each script. For instance, the devnagari script is characterised by the dominance of horizontal lines, whereas Malayalam script is characterised by the dominance of vertical lines and curves. While Roman(English) script is a good mix of linear and curved subpatterns. On the other hand, energy is distributed more or less evenly amongst different orientations for scripts which have curved patterns, like Malayalam and Gurumukhi. Here, in the proposed framework, features are learned for each script from prior knowledge(from the training data set) of target script classes. These extracted features contain finer and sufficient discriminating details of the scripts. A script class is separated from other script class by exploiting its unique features.

2.2 Feature Extraction

The energy based features are used corresponding to each script class. We use Empirical mode decomposition to capture the oriented local energy . As discussed earlier, the constructing subpatterns of each script are composed of linear/ curvilinear subpatterns. Therefore any script class can be distinguished from other based on the structural subpattern behavior.

From the prior observation(having the knowledge about structural subpatterns) of each script, we have empirically determined the angular orientation.

We have computed the oriented local energy features for the discrimination between script class. Also in Table 2, we have defined the average energy and the variance corresponding to each script class. While testing a document or defining a script class to a document, the energy corresponding to the test document is compared with the average energy corresponding to each class. The variance for each script class is also computed. The projections are taken using Radon transform at different orientations for each script class between angle 0° to 90°. The transformed script images are decomposed using EMD(Empirical mode decomposition) to analyze local characteristics and a finite set of IMF(s) are obtained. We have considered first four IMF(s) in our experiments.Since, all the Intrinsic mode functions(IMF) obtained does not contain sufficient energy in that orientaion and can not be used as discriminating feature. We have empirically determined the IMF in which the energy distribution is maximum for that script class and moreover the script class is distinguishable from other script class. The angular orientation at which Radon transform is computed and the IMF compositly represents the distinguishing feature vector for each script class.

Table 1 shows the feature vectors selected for the script classes considered for our problem.

Features Extracted for each script class			
Feat_vector	Script class	angle of orient.	IMF
FV1	Devnagari	90°	IMF1
FV2	Malayalam	90°	IMF3
FV3	Gurumukhi	$0 - 10^\circ$	IMF1
FV4	Roman	0°	IMF4
FV5	Bangla	90°	IMF2

Table 1: Feature vectors defined for script classes

The average energy and the variance corresponding to each feature vector is shown below in Table 2.

Average energy and Variance for each Feat_vect		
Feat_vector	Average energy	Variance
FV1	0.0338	$6.8798 \exp -04$
FV2	0.0243	2.5480e-04
FV3	0.0633	0.0057
FV4	0.0437	0.0028
FV5	0.0479	$8.1585 \exp -04$

Table 2: Average energy and Variance

2.3 Feature Selection

There are many potential benefits to feature selection like facilitating data visualization, data understanding, reducing training and utilization times and improving accuracy. Feature selection is selecting the most relevant variables, is usually suboptimal for building a predictor, particularly if the variables are redundant. Here, we have used the feature selection for selecting appropriate and relevant features for the problem at hand. The angle of orientation chosen for computing radon transform is totally dependent on data visualization and understanding of data. We have empirically computed the angular orientation during training by having knowledge and understanding of a script class.

Next, as aforementioned, EMD method decomposes a signal into a set of components called IMF. But all the IMF(s) are not relevant

as they all are not informative. So we have to choose only the IMF that contain useful information(maximum energy) and discarding those that share similar amounts of energy. We have not used any standard method for selection of these features, as they are not helpful for our problem. Rather we have empirically found the angular orientation and IMF which can be used a distinguishing feature for individual script class.

2.4 Classifier Design

For our problem, we use to define the most similar script class to the training images as well as to testing document. We use Adaboost binary decision tree as a classifier to improve the results. This involves training of the classifier and testing the new document for different script classes.

- Train the classifier: The system extracts the textural feature of different script classes and they depend on the characteristics of specific scripts. we have performed the training based on the distance of a training sample to each class. During training, we extract the above discussed features for our dataset and use cosine similarity measure as our measure of similarity for the discrimination. This similarity measure can be computed amongst arbitrary vectors.

- Classifying a new document: Testing a new document, we compute the features in the same way as described above. Cosine similarity measure is used for the comparison with other script classes in the dataset and the most similar script class is assigned to the test document. To improve the performance of classification, we use Adaboost binary decision tree.

2.5 Empirical Mode Decomposition

In this paper, we utilize the aspect of decomposing a signal into IMFs for analyzing nonstationary and nonlinear time series data developed by Huang et al. [3]. The intrinsic decomposition nature can adaptively decompose images into a series of modes representing different local features of images. The decomposition is based on the local characteristic of time scale of the data.

- Pre-requisites
  - the number of external and the number of zero crossing must either equal or differ at most by one;
  - should be symmetric with respect to local zero mean.

  With these two prerequisites, an IMF can be represented with a meaningful instantaneous frequency.

- Sifting Procedure
  From a given signal *x(t)*, we extract IMFs using sifting process satisfying the above defined conditions.

  1. Identify all extremas
  2. interpolate between minima with respective maxima ending up with $L_{min}(t)$ respective $L_{max}(t)$
  3. compute the mean $m(t) = (L_{min}(t) + L_{max}(t))/2$
  4. extract the details $d(t) = x(t) - m(t)$
  5. iterate the same process on the residual $m(t)$

  Once this is achieved, the detail is referred to as an Intrinsic Mode Function (IMF), the corresponding residual **m(t)** is computed and step 5 applies.

- Stopping Criterian
  If we go for sifting beyond a limit, we will get IMFs as frequency modulated signal, but constant in amplitude. So standard deviation is computed as stopping equation to stop sifting. The equation is:

$$x(t) = \sum_{i=1}^{m} IMF_i + r_m(t) \quad (1)$$

  where m is the number of IMFs obtained for a given signal and $r_m(t)$ is final residual.

3. RESULTS AND DISCUSSION

We have used 300 samples of each script class and a total of 1500 images database have been used for out experiments. These experiments have been done vice a versa for each pair of script class classification. We have performed the experiments on training and testing dataset vice a versa. So the accuracy of the experiments shown below is the average accuracy. As far as the feature extracted for script classes under consideration is concerned, it depends on the ratio of linear,curvical and curvilinear textural behavior of the subpatterns. For example the ratio of curvical and linear subpatterns are more in Devnagari and Roman than Gurumukhi and Bangla. The script classes Gurumukhi and Bangla contains more cuvilinear subpatterns than linear. While if Malyalam is considered, it contains more cuvical subpatterns than Gurumukhi and Bangla even. So the angle of orientation chosen strictly depends on ratio of the structural subpattern behavior. While selecteing the feature vector for any script class the structural behavior of must be analyzed.

Table 3 shows the average classification accuracy evaluated of all the script classes.

Classification accuracy achieved					
Script	Script classes				
	Dev	Mal	Guru	Roman	Bangla
Devnagari	97%	3%	-	-	-
Malyalam	4%	96%	-	-	-
Devnagari	97%	-	-	3%	-
Roman	5%	-	-	95%	-
Malayalam	-	94%	-	6%	-
Roman	-	3%	-	97%	-
Devnagari	92%	-	-	8%	-
Gurumukhi	10%	-	90%	-	-
Bangla	-	-	4%	-	96%
Gurumukhi	-	-	92%	-	8%

4. EXPERIMENTAL OBSERVATION

We have applied the proposed script identification(classification) method on OCR database document images. We have used for our experimentation 300 documents of each script, so dataset is composed of total 1500 document images. 50% of the images are used for training and rest 50% for testing of each script. We have used 512 by 512 document images for extracting features. The average number of characters in 512 by 512 document image is approx. 1500 to 2000. The proposed method gives better performance on document image with minimum 256 by 256 size , because it can capture sufficient distribution of pixel intensity according to strokes of each script class. We have done the experimentation on interchanging the traing and testing dataset. So the performance shown above in (Sec. 3) is the average performance of both experiments done on different document images of the dataset.

5. CONCLUSION AND FUTURE WORK

This method is developed for identification(classification) of multiplt script classes individually. The strong potential of the presented work here is its application and accuracy. The EMD(Empirical Based Decomposition) based approach has been applied in various applications but here is the first of its kind for script identification. The proposed method is computationally less time consuming, hence for script specific identification applications, it will be highly effective as compared to other expensive feature extraction and classification methods. Also the performance of the method is reasonably high. In future, We can extend this work for more number of scripts and even for different script classes. We can also apply it for text/image separation. As the energy distribution of images are evenly scattered at diiferent mode functions but in text it will always be high and oriented than images. Also the proposed method can be extended to multiple scales for script identification.

Acknowledgment

The authors are grateful to Prof. S.D Joshi, Dept. of Electrical Engg., IIT, Delhi for their helpful discussion and encouragement during this work.

6. REFERENCES

[1] D. Ghosh, T. Dube, and S. A.P. Script Recognition - A Review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32:2142 – 2161, 2010.

[2] L. Hochberg, L. Kerns, P. Kelly, and T. Thomas. Automatic Script Identification from images using Cluster-based Templates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19:176–181, 1997.

[3] H. Huang and J. Pan. Speech pitch determination based on hilbert-huang transform. *Signal Processing*.

[4] G. Joshi, S. Garg, and J. Sivaswamy. A generalized framework for script identification. *Int. J. on Document Analysis and Recognition*, page 55âĂŞ68, 2007.

[5] L. LAM, J. DING, and C. SUEN. Differentiating between Oriental and European Scripts by Statistical Features. *International Journal of Pattern Recognition and Artificial Intelligence*, pages 63–79, 1998.

[6] U. Pal and B. Chaudhuri. Script Line Separation from Indian Multi-Script Documents. In *Int. Conf. Document Analysis and Recognition*, pages 406–409, 1999.

[7] T. Tan. Rotation invariant texture features and their use in automatic script identification. *IEEE Trans. Pattern Analysis and Machine Intelligence*, pages 751–756, 1998.

[8] B. Waked, S. Bergler, C. Suen, S. Khoury, and C. Y. S. S. Khoury. Skew Detection, page segmentation, and script classification of printed document images. pages 4470–4475, 1998.

[9] N. Wang, L. Lam, and C. Y. Suen. Noise tolerant script identification of printed oriental and english documents using a downgraded pixel density feature. *Pattern Recognition, International Conference on*, pages 2037–2040, 2010.