



# Off-line Bangla handwritten word recognition: a holistic approach

Showmik Bhowmik<sup>1</sup> · Samir Malakar<sup>2</sup>  · Ram Sarkar<sup>1</sup> · Subhadip Basu<sup>1</sup> · Mahantapas Kundu<sup>1</sup> · Mita Nasipuri<sup>1</sup>

Received: 9 September 2017 / Accepted: 21 February 2018 / Published online: 3 March 2018  
© The Natural Computing Applications Forum 2018

## Abstract

Due to the cursive nature, segmentation of handwritten Bangla words into characters and also recognition of the same sometimes become a very challenging problem to the researchers. Presence of comparatively large character set along with modifiers, ascendants, descendants, and compound characters makes the segmentation task more complex. As holistic method avoids such character-level segmentation, it is generally useful for the recognition of words written in any such complex scripts. In the present work, a holistic handwritten word recognition method is developed using a feature descriptor, designed by combining different Elliptical, Tetragonal and Vertical pixel density histogram-based features. Recognition process is carried out separately using two classifiers, *namely* multi-layer perceptron (MLP) and support vector machine (SVM). For evaluation of the proposed method, a database of 18,000 handwritten Bangla word images, having 120 word classes, is prepared. The proposed system performs comparatively better with SVM than MLP for the prepared dataset. It has achieved 83.64% accuracy at best case and 79.38% accuracy on an average using fivefold cross-validation. The current method has also outperformed some recently reported holistic word recognition technique tested on the developed dataset. In addition to that the database, prepared in this work, is made freely available to fill the absence of a publicly available standard database for holistic Bangla word recognition.

**Keywords** Holistic word recognition · Handwritten word · Tetragonal feature · Elliptical feature · Vertical pixel density histogram-based feature · Bangla script

## 1 Introduction

Automatic recognition of handwritten text is one of the most popular areas of research in the domain of document image processing [1, 2]. The reason of its popularity lies in its wide range of applications in human society which include postal automation [3, 4], bank check processing [5, 6], form processing [7, 8], etc. Major difficulty in recognizing the handwritten text is mainly due to the varying writing styles of individuals. Even the script in which the text is written can pose additional challenges. For example, some Indic scripts like Devanagari and Bangla comprise a

considerably larger character set in comparison with Roman/Latin script. Chinese, Japanese and Korean scripts also have large character sets but in these scripts characters appear isolated in the text, whereas Indic scripts are very often written in cursive manner. Therefore, development of a comprehensive and accurate handwritten text recognition system in Indic script is difficult and needs more attention from the researchers [9].

In the literature, plenty of work can be found for the recognition of words written in Arabic [10–12], Chinese/Japanese [13–15] and Roman [16–18] scripts. But in comparison with that very few attempts have been made for the recognition of words written in Bangla script. With more than 200 million speakers, Bangla is the seventh most spoken language in the world [19]. It is also the second most popular official language (out of 23 official languages) in India and the national language of Bangladesh. Besides Bangla language, Bangla script is also used to write other languages like Assamese and Manipuri. Although a significant number of work have been reported

---

✉ Samir Malakar  
samirmalakar@ieee.org

<sup>1</sup> Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

<sup>2</sup> Department of Computer Science, Asutosh College, Kolkata, India

Suman Kumar Bera, Radib Kar, Souvik Saha, Akash Chakrabarty, Sagnik Lahiri, Samir Malakar and Ram Sarkar\*

# A One-Pass Approach for Slope and Slant Estimation of Tri-Script Handwritten Words

<https://doi.org/10.1515/jisys-2018-0105>

Received February 21, 2018; previously published online June 27, 2018.

**Abstract:** Handwritten words can never complement printed words because the former are mostly written in either skewed or slanted form or in both. This very nature of handwriting adds a huge overhead when converting word images into machine-editable format through an optical character recognition system. Therefore, slope and slant corrections are considered as the fundamental pre-processing tasks in handwritten word recognition. For solving this, researchers have followed a two-pass approach where the slope of the word is corrected first and then slant correction is carried out subsequently, thus making the system computationally expensive. To address this issue, we propose a novel one-pass method, based on fitting an oblique ellipse over the word images, to estimate both the slope and slant angles of the same. Furthermore, we have developed three databases considering word images of three popular scripts used in India, namely Bangla, Devanagari, and Roman, along with ground truth information. The experimental results revealed the effectiveness of the proposed method over some state-of-the-art methods used for the aforementioned problem.

**Keywords:** Slope; slant; handwritten word; oblique ellipse; eigenvector.

## 1 Introduction

Though the world is rapidly moving towards the electronic era from the traditional manual processing systems, paper documents are still in use for various applications. Digitalization of such documents can bridge the gap between past and present technologies. However, the digitalization of such handwritten documents may become useless if the existing optical character recognition (OCR) system does not convert the corresponding documents into machine-editable form properly. Generally, the task of an OCR system becomes more complex not only for the varying styles of writing but also the skewedness of transcriptions. In handwritten documents, slope and slant are inevitably introduced depending on various factors such as writing style, writing speed, or even the mood of the writer. Such complexities make the OCR process more challenging and result in poor recognition accuracy. Hence, a necessity arises to normalize such documents to an acceptable level so that the recognition system yields optimal outcome.

Slope and slant corrections are the basic pre-processing steps that have been addressed by various researchers in the last few decades. Slope-correction algorithms first measure the angle between the horizontal axis and the line along which the word image is aligned; then, the de-sloped image is generated by rotating the original image inversely according to the slope angle. On the other hand, slant-correction algorithms, in general, first find out the angle between the vertical axis and the most dominant vertical stroke of the slope corrected word, thereafter shearing the image by that angle to provide its de-slanted form.

Various attempts have been taken for page-level skew or text-line level slant correction, where a limited number of works are found for the same at the word level. We present a brief description of recent works related to our problem. Hough transform [8] is one of the most popular techniques used for correcting the

---

\*Corresponding author: Ram Sarkar, Department of Computer Science and Engineering, Jadavpur University, Kolkata, India, e-mail: raamsarkar@gmail.com

Suman Kumar Bera, Radib Kar, Souvik Saha, Akash Chakrabarty and Sagnik Lahiri: Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

Samir Malakar: Department of Computer Science, Asutosh College, Kolkata, India



Samir Malakar\*, Manosij Ghosh, Ram Sarkar and Mita Nasipuri

# Development of a Two-Stage Segmentation-Based Word Searching Method for Handwritten Document Images

<https://doi.org/10.1515/jisys-2017-0384>

Received July 30, 2017; previously published online July 4, 2018.

**Abstract:** Word searching or keyword spotting is an important research problem in the domain of document image processing. The solution to the said problem for handwritten documents is more challenging than for printed ones. In this work, a two-stage word searching schema is introduced. In the first stage, all the irrelevant words with respect to a search word are filtered out from the document page image. This is carried out using a zonal feature vector, called pre-selection feature vector, along with a rule-based binary classification method. In the next step, a holistic word recognition paradigm is used to confirm a pre-selected word as search word. To accomplish this, a modified histogram of oriented gradients-based feature descriptor is combined with a topological feature vector. This method is experimented on a QUWI English database, which is freely available through the International Conference on Document Analysis and Recognition 2015 competition entitled “Writer Identification and Gender Classification.” This technique not only provides good retrieval performance in terms of recall, precision, and F-measure scores, but it also outperforms some state-of-the-art methods.

**Keywords:** Word searching, HOG feature, topological feature, holistic word recognition, handwritten documents, QUWI database.

## 1 Introduction

Handwriting is an age-old and standard way of communication. Before the invention of modern technologies for communication, like printing press, typewriter, and email, communication was predominantly made via handwriting. Even in this era of advanced technology, a large number of people still prefer the ancient way of communication, which leads to an exponential growth of handwritten documents. Due to mismanagement, the quality of the document degrades with each passing day, which implies that preservation of documents is a pressing need. In addition, manual searching for some important information from a huge repository is time consuming and at times also error prone. Sometimes, due to lack of proper maintenance, some important document(s) may be misplaced or lost. All these issues are posing genuine problems for proper storage and maintenance of handwritten document images. Apart from these facts, advanced office automation demands mechanized storage, manipulation, and retrieval of documents in electronic format, i.e. handwritten documents need to be managed properly. The undebatable solution of this is to convert document images into an electronic form and then process the same with an optical character recognition (OCR) engine [4]. However, current handwritten OCR engines work poorly for large lexicon sizes [7]. Therefore, the alternative solution is to keep the documents in well-indexed digital form with appropriate tagging. One of the possible ways of tagging is keyword-based document indexing, which is an enriched research problem. It has a large number of real-life applications [13, 15, 25].

---

\*Corresponding author: Samir Malakar, Department of Computer Science, Asutosh College, Kolkata, India, e-mail: malakarsamir@gmail.com

Manosij Ghosh, Ram Sarkar and Mita Nasipuri: Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

# A GA based hierarchical feature selection approach for handwritten word recognition

Samir Malakar<sup>1</sup>  · Manosij Ghosh<sup>2</sup> · Showmik Bhowmik<sup>2</sup> · Ram Sarkar<sup>2</sup> · Mita Nasipuri<sup>2</sup>

Received: 19 August 2017 / Accepted: 18 December 2018 / Published online: 1 January 2019  
© Springer-Verlag London Ltd., part of Springer Nature 2019

## Abstract

Feature selection plays a key role in reducing the dimensionality of a feature vector by discarding redundant and irrelevant ones. In this paper, a Genetic Algorithm-based hierarchical feature selection (HFS) model has been designed to optimize the local and global features extracted from each of the handwritten word images under consideration. In this context, two recently developed feature descriptors based on *shape* and *texture* of the word images have been taken into account. Experimentation is conducted on an in-house dataset of 12,000 handwritten word samples written in Bangla script. This database comprises names of 80 popular cities of West Bengal, a state of India. Proposed model not only reduces the feature dimension by nearly 28%, but also enhances the performance of the handwritten word recognition (HWR) technique by 1.28% over the recognition performance obtained with unreduced feature set. Moreover, the proposed HFS-based HWR system performs better in comparison with some recently developed methods on the present dataset.

**Keywords** Hierarchical feature selection · Genetic Algorithm · Handwritten city name · Bangla script · Elliptical feature · Gradient-based feature

## 1 Introduction

In machine learning, feature selection (FS), also known as variable selection, is the process of selecting a subset of relevant features which can be used to construct a better model for any pattern classification problem. This implies FS optimizes the original feature vector through identification and removal of the redundant and/or irrelevant ones.

This optimization has many potential benefits like facilitating better data visualization and understanding of features in feature space, shrinking the storage requirement, reducing training time of a classifier and defying the curse of dimensionality to improve prediction performance.

Researchers from all over the world design various kinds of features, such as *shape-based*, *texture-based*, topological or their combinations to handle several pattern recognition problems. Such practices not only result in escalation of feature dimension, but also in most cases, fail to produce desired outcome. The key reason behind this is that generation of extra features does not always ensure their relevance in representing the patterns in the feature space or their complementary nature with respect to existing ones.

One of the possible solutions to this problem is to identify an optimal feature set which can efficiently represent the pattern classes, present in a given dataset. In this context, FS methods play an important role by reducing the feature dimension through elimination of the noisy and/or irrelevant features. As the selection of an optimal feature set is a NP-hard problem, several greedy approaches like branch and bound [1, 2], sequential search [3] and meta-

---

✉ Samir Malakar  
malakarsamir@gmail.com; samirmalakar@ieee.org

Manosij Ghosh  
manosij1996@gmail.com

Showmik Bhowmik  
showmik.cse@gmail.com

Ram Sarkar  
raamsarkar@gmail.com

Mita Nasipuri  
mitanasipuri@gmail.com

<sup>1</sup> Department of Computer Science, Asutosh College, Kolkata, India

<sup>2</sup> Department of Computer Science and Engineering, Jadavpur University, Kolkata, India



# Handwritten word recognition using lottery ticket hypothesis based pruned CNN model: a new benchmark on CMATERdb2.1.2

Samir Malakar<sup>1</sup> · Sayantan Paul<sup>2</sup> · Soumyadeep Kundu<sup>2</sup> · Showmik Bhowmik<sup>3</sup> · Ram Sarkar<sup>2</sup> · Mita Nasipuri<sup>2</sup>

Received: 16 July 2019 / Accepted: 19 March 2020 / Published online: 3 April 2020  
© Springer-Verlag London Ltd., part of Springer Nature 2020

## Abstract

Handwritten word recognition, a classical pattern recognition problem, converts a word image into its machine editable form. Mainly two basic approaches are followed to solve this problem, one is segmentation-based and the other is holistic. A number of research attempts have shown that the holistic approach performs better than its counterpart when the lexicon is predefined, fixed and small in size. Relying on this, initial benchmark recognition accuracy on CMATERdb2.1.2, a publicly available database consists of handwritten city names in Bangla, was reported following a holistic word recognition protocol. In the present work, we have followed the same trend to recognize the word samples of the said database and set a new benchmark recognition accuracy. A sparse convolutional neural network (CNN)-based model which is a low-cost trainable model has been developed for this. We have relied on a recently proposed hypothesis, known as lottery ticket hypothesis for pruning the layers of CNN model methodically, and derived a low-resource model having much less number of training parameters. This model competently surpasses the previously reported recognition accuracy on the said database by a significant margin with an axed training cost.

**Keywords** Handwritten word recognition · CNN model · Lottery ticket hypothesis · Bangla script · CMATERdb2.1.2

## 1 Introduction

The process of handwritten word recognition (HWR) aims to convert a handwritten word image into its machine editable form generally in Unicode representation. Variations in writing style of different writers make HWR a challenging task. Writing style of the writers may vary based on their sex, age, profession, educational qualification and mood. In addition to these, the type and quality of writing medium increase the complexity of the recognition task, i.e. words written on paper using pen/pencil in offline mode are more complex to recognize than writing on tablet, mobile phone, etc., using stylus or digital pen in online mode [1].

Irrespective of sample collection mode, HWR systems follow either segmentation-based approach or holistic approach while recognizing a given word. In segmentation-based approach, firstly word images are segmented into characters by using some segmentation mechanism [2]. Then, a classification model is used to recognize each of the segmented shapes, and finally reconstruction of the

✉ Samir Malakar  
samirmalakar@ieee.org; malakarsamir@gmail.com

Sayantan Paul  
sayantanpaul98@gmail.com

Soumyadeep Kundu  
soumyadeep1497@gmail.com

Showmik Bhowmik  
showmik.cse@gmail.com

Ram Sarkar  
raamsarkar@gmail.com

Mita Nasipuri  
mitanasipuri@gmail.com

<sup>1</sup> Department of Computer Science, Asutosh College, Kolkata, India

<sup>2</sup> Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

<sup>3</sup> Department of Computer Science and Engineering, Ghani Khan Choudhury Institute of Engineering and Technology, Malda, India

# Feature Selection for Handwritten Word Recognition Using Memetic Algorithm



Manosij Ghosh, Samir Malakar, Showmik Bhowmik, Ram Sarkar  
and Mita Nasipuri

**Abstract** Nowadays, feature selection is considered as a de facto standard in the field of pattern recognition where high-dimensional feature attributes are used. The main purpose of any feature selection algorithm is to reduce the dimensionality of the input feature vector while improving the classification ability. Here, a Memetic Algorithm (MA)-based wrapper–filter feature selection method is applied for the recognition of handwritten word images in segmentation-free approach. In this context, two state-of-the-art feature vectors describing texture and shape of the word images, respectively, are considered for feature dimension reduction. Experimentation is conducted on handwritten Bangla word samples comprising 50 popular city names of West Bengal, a state of India. Final results confirm that for the said recognition problem, subset of features selected by MA produces increased recognition accuracy than the individual feature vector or their combination when applied entirely.

**Keywords** Feature selection · Memetic algorithm · Wrapper–filter method  
Handwritten word recognition · Bangla script · City name recognition

---

M. Ghosh (✉) · S. Bhowmik · R. Sarkar · M. Nasipuri  
Department of Computer Science and Engineering, Jadavpur University, Kolkata, India  
e-mail: manosij1996@gmail.com

S. Bhowmik  
e-mail: showmik.cse@gmail.com

R. Sarkar  
e-mail: raamsarkar@gmail.com

M. Nasipuri  
e-mail: mitanasipuri@gmail.com

S. Malakar  
Department of Computer Science, Asutosh College, Kolkata, India  
e-mail: malakarsamir@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
J. K. Mandal et al. (eds.), *Advances in Intelligent Computing*,  
Studies in Computational Intelligence 687,  
[https://doi.org/10.1007/978-981-10-8974-9\\_6](https://doi.org/10.1007/978-981-10-8974-9_6)



# An Advanced Particle Swarm Optimization Based Feature Selection Method for Tri-script Handwritten Digit Recognition

Suryadipto Sarkar<sup>1</sup>(✉), Manosij Ghosh<sup>2</sup>, Agneet Chatterjee<sup>2</sup>,  
Samir Malakar<sup>3</sup>, and Ram Sarkar<sup>2</sup>

<sup>1</sup> Department of Electronics and Communication Engineering,  
Manipal Institute of Technology, Manipal, Karnataka, India  
ssarkarmanipal@gmail.com

<sup>2</sup> Department of Computer Science and Engineering, Jadavpur University,  
Kolkata, India  
manosij1996@gmail.com, agneet257@gmail.com,  
raamsarkar@gmail.com

<sup>3</sup> Department of Computer Science, Asutosh College, Kolkata, India  
malakarsamir@gmail.com

**Abstract.** Handwritten digit recognition is a well-studied pattern recognition problem. Most of the techniques, reported in the literature, have been concentrated on designing several feature vectors which represent the digits in a better way. But, at times, such attempt not only increases the dimensionality of extracted feature vector but also suffers from having irrelevant and/or redundant features. To address this, in the present work, a recently introduced Particle Swarm Optimization (PSO) based feature selection method has been applied with suitable modifications. For the course of this experiment, we have confined ourselves to a newly employed feature vector for handwritten digit recognition, namely DAISY feature descriptor. The proposed feature selection method is tested on three handwritten digit databases written in Bangla, Devanagari and Roman scripts. The experimental results show that significant amount of feature dimension is reduced without compromising on recognition accuracy. Comparison of the present feature selection method with two of its ancestors also reveals that the present method outperforms the others.

**Keywords:** Handwritten digit recognition · DAISY descriptor · Feature selection · Binary PSO · Bangla · Devanagari · Roman

## 1 Introduction

High variation in handwriting of different individuals is commonplace. This variation is also observed for a single individual when s/he writes the same text at different times. As a result, recognition of unconstrained handwriting becomes challenging and is still considered an open research topic to the document analysis community. These facts are equally applicable for handwritten digit recognition, which is considered as one of the classical pattern recognition problems. The reason behind this is the spectrum of utility



# Textual Content Retrieval from Filled-in Form Images

Soulib Ghosh<sup>1</sup>, Rajdeep Bhattacharya<sup>1</sup>(✉), Sandipan Majhi<sup>1</sup>,  
Showmik Bhowmik<sup>1</sup>, Samir Malakar<sup>2</sup>, and Ram Sarkar<sup>1</sup>

<sup>1</sup> Department of Computer Science and Engineering, Jadavpur University,  
Kolkata, India

ghoshsoulib@gmail.com, rajdeep.cse17@gmail.com,  
sandipan.majhi.email@gmail.com, showmik.  
cse@gmail.com, raamsarkar@gmail.com

<sup>2</sup> Department of Computer Science, Asutosh College, Kolkata, India  
malakarsamir@gmail.com

**Abstract.** Form processing refers to the process of extraction of information from filled-in forms. In this work, we have addressed three very crucial challenges of a form processing system, namely touching component separation, text non-text separation and handwritten-printed text separation. The proposed method is evaluated on a database having 50 filled-in forms written in Bangla, collected during an essay competition in a school. The experimental results are promising.

**Keywords:** Form processing · Text non-text separation ·  
Handwritten-printed text separation · Touching component separation ·  
Bangla text

## 1 Introduction

Form processing refers to the process of extracting the textual information present in a filled-in form. Forms have been preferred for information collection in various departments such as railway, bank, educational organization, administrative office, etc. Forms can be divided into various categories such as orders, applications, claims and survey forms. Huge volume of such forms generated in every department makes manual processing tedious. Thus, development of an automated form processing system becomes a pressing need.

Processing of filled-in forms, however, has many challenges such as the diversity that exists in the type of the data (e.g. numerals, alphabet, etc.), occurrence of various non-text elements (e.g. tables, lines of different types, logs, etc.) and presence of both printed and handwritten text. In addition to these, two more critical problems are there, one of which appears at the text extraction level which is occurrence of touching components, and the other appears at the text recognition level due to the complexity of handwritten text. In this work, a system is developed for the extraction of text present in a filled-in form, where we have addressed three major issues namely, touching component separation, text non-text separation, and handwritten-printed text separation.