A Novel Approach to Script Separation

Ranjith Kumar, Vamsi Chaitanya and C. V. Jawahar
Centre for Visual Information Technology
International Institute of Information Technology
Gachibowli, Hyderabad 500 019, INDIA
jawahar@iiit.net

Abstract

This paper describes a new approach for script separation. A character level script separation scheme is combined with a Viterbi algorithm to get an optimal sequence of scripts which could generate such a text. This method complements the popular approaches for script separation at paragraph level using texture features or at line level using structural features.

1. Introduction

Most character recognition systems assume that the script of the document (or Image block) is known prior to the processing. If the document image contains characters from multiple scripts or languages, the recognition problem becomes difficult. Once the scripts are identified, document understanding systems could use powerful algorithms that incorporate script-specific and contextual information for higher performance. This problem of script separation achieves special emphasis in the Indian multilingual context. A simple classification of script (say as Han-based or Latin-based) does not help the recognition process enough. There can be more than one language which use the same or highly similar script and the corresponding language models could be considerably different.

There have been many initiatives in distinguishing Latin, Han and Arabic scripts in literature [7], [3]. One of the approaches [1] is to extract the attributes of several connected components (say a paragraph or line) and then compare with templates representing individual scripts. Similar approaches by computing the structural or geometric features for lines of text is reported for many scripts [9]. A second approach [8] uses the fact that a script has a distinctive visual appearance and extracts the textural features of a script for identification.

In a typical document, a paragraph need not be of a homogeneous script. In many Indian official documents, at least three languages (English, national and regional) are often observed. They are also interleaved in such a way that script separation algorithms at line or paragraph level is not accurate enough. However, most of the consecutive words are of same script and script changes only once in a while or depending on some probabilities, which could be modelled. This paper aims to achieve script separation by exploiting this idea. This approach complements the work done at paragraph/page or line/sentence level and could be integrated with the structural feature-based approaches for complete solution to the script separation problem.

Script separation for Indian language documents is extensively investigated by Pal and Chaudhuri [4, 3], primarily using structural features. They [9, 3] conducted script separation studies very effectively for many pairs of languages most of the time at line level and rarely at paragraph or word level. This paper does not intend to solve the script separation between any specific pair or among Indian languages. Instead, this paper, reveals a complementary direction to the script separation research without using structural or textural features. We attempt script separation at character level with errors and use a Viterbi algorithm later to correct these errors.

Section 2 discusses a simple mechanism for recognition of script at character level using a Neural Network. The performance at this character level is improved by considering the script separation as an analysis of a sequence of outputs from a set of states. A Viterbi-algorithm based optimal script identification is explained in Section 3 under various situations. A discussion on the prospects and limitations of this approach is explained in Section 4. Section 5 describes the concluding remarks.

2. Script Separation at Character Level

Most of the script recognition algorithms work at the paragraph or line level. It is observed that with smaller units for recognition, problem becomes more and more complex, while they remain more useful. The problem of script separation at character level is particularly difficult for Indian languages, since there are many similar characters in different scripts, with different ISCII or UNICODE representa-

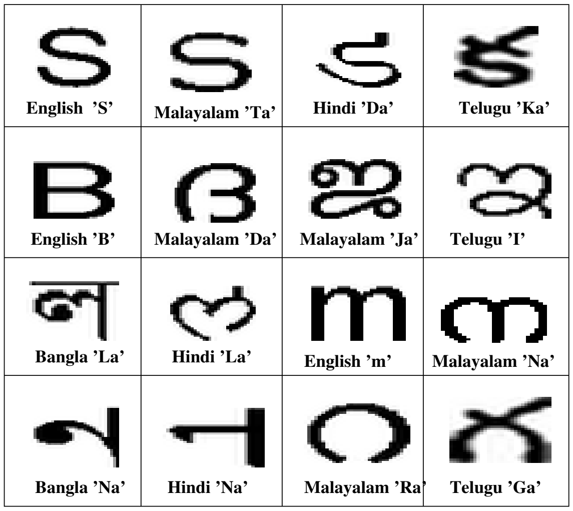


Figure 1: Why character level script separation is difficult

tions. These characters can be identical or highly similar. Figure 1 shows a set of such characters. The similar looking set of characters come from different script making the script separation problem practically impossible to solve at character level, even with the help of an excellent classifier. Our study concludes that script separation can be done with limited errors. However, knowledge about the structure of the document and common script change behaviour can correct most of these errors present at this stage, as we demonstrate in the next section. Another problem with the script separation at character level is the presence of large intra-class variations. This makes the classifier boundaries highly complex and practically difficult to estimate. Since the intra-class variations are quite high for a script, we decided to employ a Neural network based classifier for the recognition.

Multi-Layer Perceptron [6]: Multilayer perceptrons (MLP) are popular to solve supervised pattern recognition problems. This network, as a classifier, can be used to approximate highly non-linear decision boundaries. Networks are trained using backpropagation algorithm. In practice, they are often shown to get overtrained or getting stuck in local minima if enough care is not taken while training. Even with these limitation, they are still considered to be very useful for pattern classification problems. We have considered a three-layer network for the classification. A set of script-labelled characters are used for training. Input layer process normalised character images and target is the script label. With training, network leans to approximate the image to scriptID mapping. Part of the data set used are from experiments reported in [2]. The character images were scaled into a fixed size of 30 × 30 and a multilayar perceptron with backpropagation algorithm is used for the training. Network has 900 input nodes, and outputs based on the number of scripts used in the experiment. One hidden layer with approximately 1000 neurons is employed.

Experiments: Extensive experimentation has been carried out to learn the script separation behaviour at the character level. Sample images from four Indian languages – Hindi, Telugu, Bangla and Malayalam – were considered along with the English characters. Structural features like presence of *sirorekha* are important clues for separation of many Indian language pair. Note that we have not employed this yet in this formulation. Instead, sample images which are used at the OCR-level (*sirorekha* removed zone-wise-segmented components) are employed for the script separation. For the experimentations, around 18000 samples were taken for the training and another 18000 samples obtained from the same population is considered for testing. On many pairs like Telugu-Hindi, simple neural network classifier provided excellent results, where the training and testing data have same fonts/characteristics and variability in samples are limited. Results of the order of 98% were obtained for many pairs. However, for Bangla-Hindi pair results were rather low in the tune of 72% . On a highly diverse sample set, performance of the script separation came down drastically. Only an average performance of around 71% was obtained. For example with training and testing data from Arial and Kartika for English and Malayalam scripts respectively, we could obtain an accuracy of 98% for English and 94.63% for Malayalam. However when tested with a mixture of Arial(bold,talic) Verdana,Times for English and Kartika(bold,normal,italic) for Malayalam, the accuracies came down to 89.19% and 66.69% respectively.

3. Script Separation using Viterbi Algorithm

Script separation at character-level is not robust, and is not enough for any applications. It is safe to assume that the script changes only at the word level. In many situations, there could be a special symbol (like line break or parenthesis) to provide possible script transition in a text. We use this information along with the evidence provided by the character-level classification to obtain script separation at word level. A simplification of the above argument will result in a simple majority based script labelling for individual words. However, in many structured documents, we have additional probabilistic information, which is exploited in our formulation.

Lets represent each script by a scriptID. Consider the text as an ordered sequence of script labels where script of each individual character is determined using the neural network

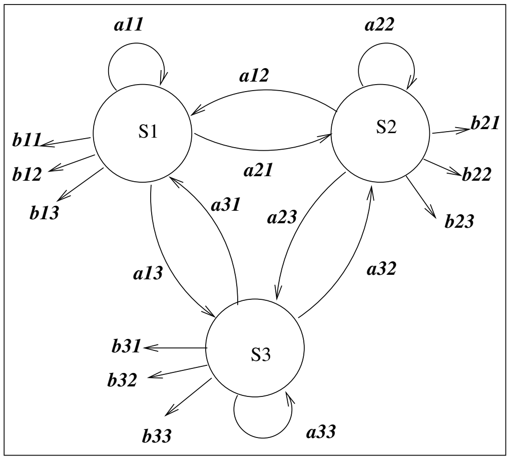


Figure 2: State Diagram Formulation of the Script Separation Problem

as in the previous section. The script changes in the sequence can be modelled as state transition process.

Consider Figure 2, where each of three states correspond to one of the scripts S_1, S_2 or S_3. At each state S_i characters can be misclassified as belonging to state S_j with the probability $b_i(S_j)$(or b_{ij}). This probability corresponds to the misclassification of character of script S_i as character of script S_j by the neural network. Let a_{ij} be the probability of transition from state S_i to S_j. Given the sequence of script labels the state sequence that maximizes the probability of the this sequence is the optimal script sequence we need. This is known as the Decoding problem of a Hidden Markov Model.

Viterbi Algorithm Viterbi Algorithm provides a dynamic programming based solution to the identification of optimal state sequences from the given set of observations. It is an inductive algorithm which at each instant keeps track of optimal state sequence for each of the given states as the intermediate state for the desired observation sequence. Out of these the one that maximizes the probability of the observation sequence is chosen. This reverses many decisions made by the neural network at the same time, the accuracy at word level improves.

We considered a long sequence of words (with no space in between them) from Malayalam and English for the experimentation. These characters were either synthetically generated (by rendering text) or by concatenating the samples described in the previous experiment. It is observed that (a) The proposed plan is effective since it uses the sequence information. (b) It corrects most of the misclassifications if the results at the character level are highly accurate. (c) In many situations, the character level recognition of script is not accurate and the proposed approach does not solve the problem very effectively.

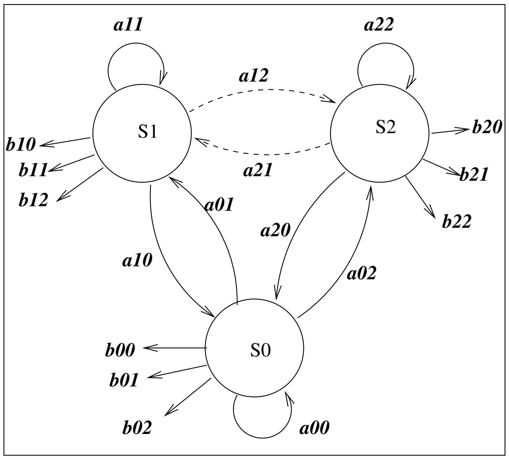


Figure 3: White Spaces and Special Characters as clue for script separation

years	എന്നും	രണ്ടായി	തിരിക്കാം.	INPUT
E M ME E	M M ME	M M M M E	M EME M E M	CHARACTER
E E E E E	M M MM	M M M M M	M MMMM M MM	AFTER VITERBI

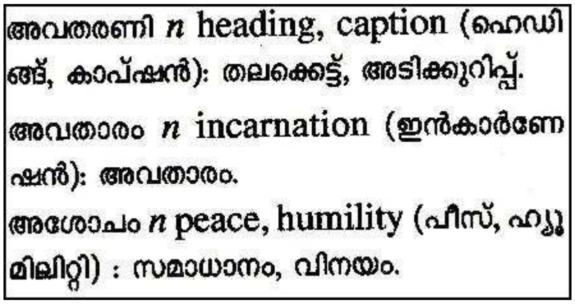
Figure 4: A sample text from an English Malayalam document and its recognised labels at character level and after Viterbi

3.1. Word Boundaries: Better Clues

Infact, in practical situations, script does not change arbitrarily. Scripts may change with paragraphs, lines or even words. But usually a word is homogeneous in script and scripts are separated by a blank space. We can use this information effectively by introducing a special state ($S0$) for special characters like white space as shown in Figure 3.

Here the transition from one script to other takes place only via a state $S0$. The state transition probabilities from one script to other is made zero. However, we have considered a small probability (order of 0.01) for the inter script transitions to handle even some unexpected situations. Here we assume that the special characters (like white spaces) are recognised correctly.

The script transition probabilities can be modelled as follows. If the scripts are not separated by special characters, transition probabilities are estimated from training sequences. In presence of special characters we use the word length statistics in the script [5]. The transition probability a_{i0} is $\frac{1}{1+w(i)}$ where $w(i)$ is the average word length in the script S_i. The probability a_{0i} may be related to the apriori probabilities of the scripts in the document.



അവതരണി *n* heading, caption (ഹെഡി
ങ്ങ്, കാപ്ഷൻ): തലക്കെട്ട്, അടിക്കുറിപ്പ്.
അവതാരം *n* incarnation (ഇൻകാർണേ
ഷൻ): അവതാരം.
അശോഷം *n* peace, humility (പീസ്, ഹ്യൂ
മിലിറ്റി) : സമാധാനം, വിനയം.

Figure 5: A sample document image from an English Malayalam dictionary

3.2. Using Probabilistic Structure

We had seen that space is an important clue in script change. There are many other important document structures which can help in pointing out the script change. Consider Figure 5, which shows a sample segment from a Malayalam-English bilingual dictionary where the script change is often associated with a parenthesis, colon or newline. This situation is possible in many multilingual documents.

Experiments: In our experimentation with structured documents, we have assumed that special characters can be reliably recognised. And all such special characters are represented together in state S_0. Many multi-script documents from the five Indian languages were considered. Most of them were from Malayalam-English combination.

We have discussed the character-level classification issues for Malayalam-English in the previous section. In such a document, we could correctly classify 151 words while 3 words were misclassified. At an average for many pairs, an accuracy of the order of 97% is acheived. The errors present were mainly on words of size two or three characters. With highly noisy data, when the script separation accuracy degrades close to $50-60\%$, output of viterbi algorithm is only around 89%.

4. Discussions

From the experiment we infer that probabilistic structure of a document simplifies its analysis greatly. A large set of documents like multilingual dictionaries, scientific literature and automatically generated documents have such structure. Modelling this structure by examining a subset of them can help us in analysis of similar documents.

This method conceptually differs with the reported methods for script separation. However, this needs character-level segmentation for many languages. The sensitivities associated with the granular-level segmentation will be shadowed by the correction process due to Viterbi Algorithm.

Further improvements can be made to the proposed approach by considering information like the second best match by the neural network, confidence factors from a recognition system etc. Another important issue in structured document is the estimation of probabilities for the Markov model. This needs a closer look.

5. Conclusions

A novel method to script separation without the help of structural and textural features is proposed in this paper. A solution to the script separation problem for a set of languages can be built by integrating the language specific information, probabilistic structure of the document, confidence values available from the character-level classification etc.

Acknowledgment

One of the authors, Vamsi Chaitanya, a summer research fellow from IIIT Calcutta wants to acknowledge the financial support provided by the JNCASR, Bangalore.

References

[1] J. Hotchberg. Automatic script identification from images using cluster-based templates. In *Third International Conference on Document Analysis and Recognition*, page 378, August 1995.

[2] C. V. Jawahar, MNSSK Pavan Kumar, and S. S. Ravikiran. A bilingual ocr system for hindi and telugu documents and its applications. In *Proc of ICDAR*, pages 403–408, 2003.

[3] U. Pal and B.B.Chaudhuri. Automatic identification of english, chinese, arabic, devnagari and bangla script line. In *Sixth International Conference on Document Analysis and Recognition (ICDAR '01)*, page 790, September 2001.

[4] U. Pal and B. B. Chaudhuri. Script line separation from Indian multi-script documents. In *Proc. Int. Conf. Document Analysis and Recognition (ICDAR)*, pages 406–409, 1999.

[5] Prakash Rao, Sushma Bendre, and Rajeev Sangal. Basic statistical analysis of corpus and cross comparison of parallel corpora. In *Proc of ICON*, pages 121–129, 2003.

[6] Simon Haykin. *Neural Networks 2nd ed.* Pearson Education, 2001.

[7] A. L. Spitz. Determination of script and language content of document images. *IEEE Transactions on PAMI*, Vol. 19(3):235–245, March 1997.

[8] T.N.Tan. Rotation invariant texture features and their use in automatic script identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(7):751–756, July 1998.

[9] U.Pal and B.B.Chaudhuri. Automatic separation of words in multi-lingual multi-script indian documents. In *Fourth International Conference on Document Analysis and Recognition*, page 576, August 1997.