2009 10th International Conference on Document Analysis and Recognition

Information Retrieval Model for Online Handwritten Script Identification

Guo Xian Tan[1, 2], Christian Viard-Gaudin[2], Alex C. Kot[1]
[1]*Nanyang Technological University, Singapore*
[2]*IRCCyN, UMR CNRS 6597, Ecole Polytechnique de l'Université de Nantes, France*
tanguoxian@pmail.ntu.edu.sg, christian.viard-gaudin@univ-nantes.fr, eackot@ntu.edu.sg

Abstract

Script identification has always been a topic of much research interest in the field of document analysis. The accurate determination of the identity of the script is paramount to many post-processing steps such as document sorting, translation and in determining the choice of linguistic resources to use for OCR or handwriting recognition. However, few works exist with regards to the identification of online handwritten scripts, partly due to the large variations and challenges innate in handwritten scripts. This paper proposes a novel approach for online handwritten script identification based on the Information Retrieval model. We attempt to identify among three script families; Arabic, Roman and Tamil scripts, which attained an average accuracy of 93.3% from our results. This signifies promising potential in utilizing Information Retrieval models for script identification.

1. Introduction

Script identification has long been the forerunner of many OCR processes as a precursor during the preprocessing stages. The identification of scripts is essential to facilitate many important further processing steps such as document sorting, indexing and translation. Another application of script identification of vital importance especially for handwritten documents in particular is the selection of a linguistic resource for text recognition. Difficulties inherent in segmenting and recognizing handwritten text due to the large variations in handwriting styles pose huge challenges. Hence, handwritten text recognizers generally make use of linguistic resources to improve the recognition results. An automatic selection of a linguistic resource for the handwriting recognition engine based on the information provided by script identification will certainly prove to be invaluable. This automatic selection of linguistic resources is especially useful in scenarios where there exist multiple scripts or languages in the same document.

In order to avoid any ambiguities between the terminologies used, we define a clear notion between script and languages here. Scripts are defined as a set of graphical representations used to express a particular writing system [1]. As seen from figure 1, it clearly illustrates that scripts are subsets belonging to a particular writing system. Languages may then make use of writing styles belonging to more than one script family, such as in the cases where the Malay language uses both the Arabic and Roman scripts or the Japanese language using all three from the Roman, Chinese and Kana scripts. Hence, there is value in performing script identification even within the same language. For instance, this allows us to distinguish between portions of the documents in the Malay language that are written in Arabic or Roman scripts.

Writing System	Scripts		Languages
Alphabet	Roman	*Hello World*	English, French, German, Indonesian, Italian, Malay, Japanese (Romanji)
	Greek	α β ε θ ω	Greek
	Cyrillic	й ж ф ю	Bashkir, Chechen, Mongolian, Russian
	Korean	안녕하세요	Korean (Hangul)
Semanto-Phonetic	Chinese	丶 一 丨 丿 乀	Chinese, Korean (Hanja), Japanese (Kanji)
Syllabic-Alphabet	Devnagari	देवनागरे	Hindi, Marathi, Nepali, Sanskrit, Sindhi
	Thai	สวัสดีปี	Thai
	Tamil	தித்திக்கும்	Tamil
Syllabary	Kana	ゥ ー ロ ン, い ろ	Japanese (katakana, hiragana)
Abjad	Arabic	فرساودار	Malay (Jawi), Morisco, Pashto Persian/Farsi, Punjabi, Sindhi
	Hebrew	וַיְהִי כָל-הָאָרֶץ	Hebrew, Judeo-Arabic, Ladino, Yiddish

Figure 1. Writing systems, script families and their languages.

Online handwritten documents not only provide information obtainable from offline digital documents, but also contain temporal information of the handwriting process [2]. Such additional information offer valuable clues as to the identities of the script.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.162

For example, certain scripts such as Arabic and Hebrew scripts adopt a right-to-left style of writing and the temporal information available in online documents allows us to differentiate such scripts apart from those scripts that use left-to-right styles of writing.

Even with this temporal advantage of online documents, most of the publications in script identification still deal with scanned documents; where they are either processing scanned images of handwritten or printed documents [3-8]. Early works by Hochberg et al. [5] and Spitz et al. [3] lay the foundation for many of the script identification works today. Spitz et al. proposed using the upward concavities found in character structures to identify between Han based scripts and the Roman script. Hochberg et al. presented a feature-based approach using connected components analysis to identify among six scripts, attaining an average accuracy of 88%. To the best of our knowledge, only an exhaustive list of works have been published in online handwritten script identification such as the works by Lee et al. [4] and Namboodiri et al. [6]. Lee et al. proposed using a hierarchical HMM approach to identify between English and Korean scripts. Namboodiri et al. made use of 11 different spatial and temporal features to identify among six different scripts, attaining an overall average accuracy of 95.5%.

The originality of our work lies in the design of a framework for online handwritten script identification that is based on Information Retrieval (IR) techniques. Our results show that high accuracies are attained for the identification of three scripts, namely Arabic, Roman and Tamil scripts using this IR-based approach. The remainder of this paper is organized as follows: Section 2 provides a discussion on the Information Retrieval (IR) model. Section 3 then goes on to describe the proposed methodology and experimental setup that makes use of this IR model in script identification. Thereafter, section 4 presents a discussion on the experimental results. Finally, conclusions and the future directions in our research are given in section 5.

2. Information Retrieval Model

Information retrieval (IR) techniques have been widely used in many applications such as writer identification, biomedical fields, web-based document searches, video retrievals and content-based image retrievals [9-11]. The general major models of IR include the Boolean model, the vector space model, the probabilistic retrieval model and the knowledge-based model [12, 13], each with their own advantages and disadvantages. The vector space model approach that was first proposed by Salton et al. [14] remains to this day one of the most dominant approaches in IR due to its relatively simple, yet effective design. The vector space model utilizes occurrence vectors in a two-pronged approach involving an indexing step and a retrieval step. In the indexing step, the set of documents is transformed into a set of occurrence vectors, the term frequency (*tf*) and inverse document frequency (*idf*). The fundamental idea of these *tf-idf* vectors is that it prescribes the degree of relevance of a particular feature depending on how frequently that feature exists in the document. Thereafter in the retrieval step, these *tf-idf* vectors of the query document are then matched with the *tf-idf* vectors of the indexed document. The document that is retrieved is the one with the most similar match between the *tf-idf* vectors. Similarly in our context of script identification, we can draw inspiration from this concept of *tf-idf* occurrence vectors to solve our problem. Hence, we have designed our script identification based on this approach.

3. System Architecture

In this paper, we propose using an IR technique to identify between three different scripts. The proposed framework can be divided into three stages according to the vector space model, which consists of the prototype building stage, the document indexing stage and the retrieval stage. The design of our script identification system is illustrated in figure 2. Details of each stage are discussed in the following subsections.

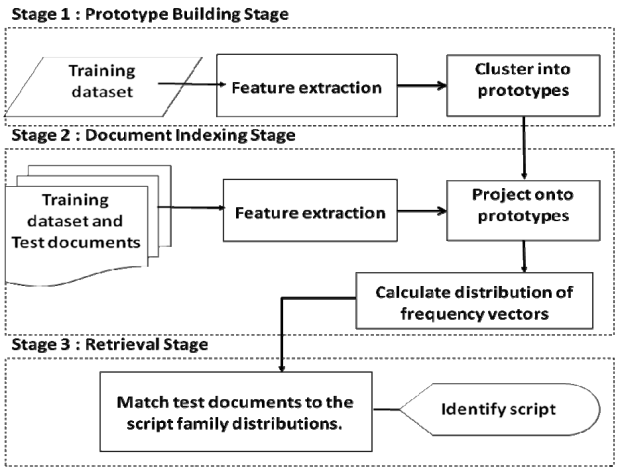

Figure 2. Block diagram of script identification.

3.1. Prototype Building Stage

The purpose of the prototype training stage is to build a set of prototypes that model the different script

families based on the extracted features. The prototypes are built and trained using the training set (discussed in details in section 4.1). Preprocessing, line extraction and the extraction of features at the line level based on the work described by Namboodiri et al. [6] is then carried out. The features selected are the horizontal and vertical interstroke direction, horizontal and vertical stroke direction, average stroke length, stroke density and the reverse direction. We have implemented only seven out of the eleven features described in [6]. This is because we do not need to identify Devanagari scripts, therefore the extraction of information related to the 'Shirorekha' is omitted in our design. The aspect ratio is also not used in this paper because this feature is only meaningful for word segments [6] and the results in this paper are reported based on line segments. Furthermore, the variance of the stroke length is not considered because we have found that this is not a very discriminating feature in our system. According to Namboodiri et al.'s results [6], this feature is the least salient among all the features they used, hence we have omitted this feature as well. The line segments extracted from the training sets are then clustered using the well-established k-means clustering algorithm [15] on a script family basis, script family by script family. This gives us a total of *3 x N* prototypes, where we have determined *N=10* experimentally.

3.2. Document Indexing Stage

The purpose of this stage is to calculate a distribution of frequency vectors for each test document to match with the distribution of frequency vectors for the script families. We have grouped all the documents belonging to the same script family in the training set into one training document *i,* in order to build the distribution for each of the three script families. With regards to the test set (refer to section 4.1), one distribution of frequency vectors is computed per test document. The computation of the distribution is achieved by mapping the extracted features from every line of a document to all of the 3 *x N* prototypes built in the previous prototype training stage. The assignment to the respective prototypes is based on a fuzzy c-means algorithm [15, 16] which uses a Cauchy kernel function as described in Eq. 1 to determine a partial membership to the respective prototypes.

$$P(p_{\varsigma,k} \mid x) = \frac{\dfrac{1}{1+\dfrac{\left(dist(p_{\varsigma,k}, x)\right)^2}{\gamma}}}{\sum_{\varsigma=1}^{S}\sum_{k'=1}^{N} \dfrac{1}{1+\dfrac{\left(dist(p_{\varsigma,k'}, x)\right)^2}{\gamma}}} \quad (1)$$

$P(p_{\varsigma,k} \mid x)$ is the probability that a given feature vector x is assigned to the k^{th} prototype $p_{\varsigma,k}$ $k \in [1, N]$ of script ς, where *S* is the number of script families and *N* is the number of prototypes used. $dist(p_{\varsigma,k}, x)$ represents the Mahalanobis distance, which takes into account the specificities of the feature space and the resulting shape of the clusters in this space. In Eq. 1, γ is a tuning parameter which is experimentally set to be *0.1*, which allows the spatial selectivity of the prototypes to be expressed.

$$tf_{\varsigma} = \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{N} P(p_{\varsigma,k} \mid x_m) \quad (2)$$

$$idf_{\varsigma} = \log \frac{\sum_{\varsigma'=1}^{S}\sum_{i=1}^{S} tf_{\varsigma',i} + \varepsilon}{\sum_{i=1}^{S} tf_{\varsigma,i} + \varepsilon} \quad (3)$$

$P(p_{\varsigma,k} \mid x_\alpha)$ is then used to calculate the term frequency (*tf*) vector, as shown in Eq. 2. Each component of the term frequency vector will be weighted with the inverse document frequency (*idf*), as shown in Eq. 3, where ε is a small value to prevent any numerical problems.

We have defined the *tf-idf* measures as shown in Eq. 2 and Eq.3. As shown in Eq. 2, the *tf* is found by summing up the contribution from all the *N* prototypes belonging to that script family. This is then repeated and normalized for all the line segments of the document, where M is the number of line segments corresponding to that document in Eq. 2. The interpretation of *tf* in our script identification scenario is that it provides an indication of how similar the document is to a particular script family. As shown in Eq. 3, the *idf* is found by only computing the training documents *i* and does not involve the test documents. The *idf can* be interpreted as how frequent features belonging to one script family is being used in other script families. For example, features not commonly present across all script families (ie. a low tf_{ς}) will give rise to a high *idf* value. This indicates that since this set of features is uniquely found only in certain script families, it will be more interesting to place

emphasis on such features by using the *idf*. Hence, *idf* only involves the distribution of script families in the training set.

3.3. Retrieval Stage

Script identification is achieved in this final retrieval stage by comparing the distribution of *tf-idf* vectors of the test document in query to that of the distribution of the three script families. We have used the minimum Chi-square distance classifier for the classification as shown in Eq. 4.

$$\text{Dist}(\text{script}_i, \text{doc}_T) = \sum_{\varsigma=1}^{S} \frac{idf_\varsigma \left(tf_{\varsigma,i} - tf_{\varsigma,T}\right)^2}{tf_{\varsigma,i} + tf_{\varsigma,T}} \quad (4)$$

4. Experimental Results

4.1. Database

Thirty handwritten documents for each of the three script families (30x3=90 documents); Arabic, Roman and Tamil scripts were collected from a total of 62 writers using a digital pen and paper. A handwritten sample from each of the three script families is shown in figure 3. The length of each document ranged from 1 to 6 lines of text. Twenty of the handwritten documents from each script family were then used for the training set. Therefore, we have 60 documents for the training set and 30 documents available in the test set in total.

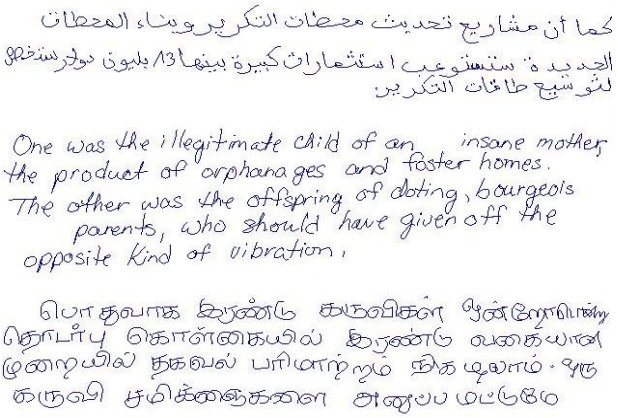

Figure 3. Sample of handwritten documents from the three script families (top: Arabic, middle: Roman, bottom: Tamil).

4.2. Discussion

An average accuracy of 93.3% was achieved on the 30 test documents while trying to identify among the three different script families. Table 1 shows the confusion matrix. Our results reveal that all the Arabic scripts have been correctly identified. Arabic scripts make use of a right-to-left style of writing and it is the only script that adopts such a unique style of writing among the three script families. There only exist few scripts such as Arabic and Hebrew scripts that make use of this writing direction. Hence such distinctive information should be fully exploited in script identification. The work by Hochberg et al. [5] could not differentiate Arabic scripts from other left-to-right scripts and only attained an accuracy of 89% for their classification of Arabic scripts because they did not utilize such information. It is also noteworthy to mention that temporal information on the writing directions can be easily derived from online handwritten documents, making them ideal for script identification.

Table 1. Confusion matrix showing the results of the three scripts.

	Classified		
True	**Arabic**	**Roman**	**Tamil**
Arabic	10	0	0
Roman	0	9	1
Tamil	0	1	9

Upon closer examination of our results, we observe that one Tamil script has been confused with the Roman scripts and one Roman script has been confused with the Tamil scripts. Both these scripts have the same writing direction and we also note that they have similar stroke length and density, which could explain the misclassifications. Comparisons with the work on the identification of handwritten Arabic and Roman scripts by Ben-Jlaiel et al. [17] show that our proposed methodology outperforms them. They attained an accuracy of 96% and 84% for identifying handwritten Arabic and Roman scripts respectively, in contrast to our improved results of 100% and 90% respectively. Another comparison that is closer to our work on online handwritten script identification [6], which we based our feature extraction method on, reveals that one Roman script was confused with their Arabic scripts and that three Devanagari scripts were confused with their Roman scripts. They managed to attain an overall accuracy of 95.5% on six different script families. We will expand our approach to a larger database containing more script families to allow for a more realistic comparison in the near future.

5. Conclusion

This paper discussed an approach that involved the use of IR techniques to build prototypes to model the script families as stochastic distributions of frequency vectors, thus providing a simple and effective method of indexing and retrieving the identity of the script. The advantage of the IR method is that it allows the frequency of occurrence of features pertaining to that script to be taken into account through the usage of the *tf-idf* combination, hence similar scripts such as Hangul and Chinese which share common features can be better differentiated using their relative contribution to the frequency vectors. Our proposed method attained an average accuracy of 93.3% based on three scripts, showing promising value in utilizing IR techniques for script identification. We will subsequently expand this technique to a larger database containing more script families and more documents. This will allow for a more realistic assessment of our script identification system. Another future area of work will address using segments at the word level for identifying the script. This will allow us to identify multi-script documents where different portions of the document can contain different scripts. This is commonly found in Japanese documents where a mixture of Japanese (Kanji) and Japanese (Hiragana, Katakana) is written in the same line. In addition, working at the word level permits us to extract more information about the identity of the script such as the aspect ratio. Typically, the aspect ratio of Tamil words is generally longer compared to the aspect ratio of other scripts such as Chinese scripts. We foresee that this can provide an interesting measure of the script identity.

Acknowledgements

This research is jointly supported by Nanyang Technological University RSS Grant in Singapore, the French Merlion Scholarship and the ANR Grant CIEL 06-TLOG-009.

Reference

[1] F. Coulmas, *Writing Systems: An Introduction to Their Linguistic Analysis*: Cambridge University Press, 2003.

[2] R. Plamondon and S. N. Srihari, "On-line and off-line handwriting recognition: A comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22, pp. 63-84, 2000.

[3] A. L. Spitz, "Determination of the script and language content of document images," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 19, pp. 235-245, 1997.

[4] J. J. Lee, J. H. Kim, and M. Nakajima, "A hierarchical HMM network-based approach for on-line recognition of multi-lingual cursive handwritings," *Ieice Transactions on Information and Systems,* vol. E81D, pp. 881-888, Aug 1998.

[5] J. Hochberg, K. Bowers, M. Cannon, and P. Kelly, "Script and language identification for handwritten document images," *International Journal on Document Analysis and Recognition,* vol. 2, pp. 45-52, 1999.

[6] A. M. Namboodiri and A. K. Jain, "Online Handwritten Script Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 26, pp. 124-130, 2004.

[7] A. Busch, W. W. Boles, and S. Sridharan, "Texture for script identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, pp. 1720-1732, 2005.

[8] G. D. Joshi, S. Garg, and J. Sivaswamy, "A generalised framework for script identification," *International Journal on Document Analysis and Recognition,* vol. 10, pp. 55-68, Nov 2007.

[9] G. X. Tan, C. Viard-Gaudin, and A. Kot, "Automatic Writer Identification Framework for Online Handwritten Documents Using Character Prototypes," *Pattern Recogn.,* 2009.

[10] W. Hsu, S. Antani, L. R. Long, L. Neve, and G. R. Thoma, "SPIRS: A Web-based image retrieval system for large biomedical databases," *International Journal of Medical Informatics,* vol. In Press, Corrected Proof, 2008.

[11] K. H. Liu, M. F. Weng, C. Y. Tseng, Y. Y. Chuang, and M. S. Chen, "Association and temporal rule mining for post-filtering of semantic concept detection in video," *IEEE Transactions on Multimedia,* vol. 10, pp. 240-251, 2008.

[12] S. M. Chen and Y. J. Horng, "Fuzzy query processing for document retrieval based on extended fuzzy concept networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,* vol. 29, pp. 96-104, 1999.

[13] T. Radecki, "Generalized Boolean methods of information retrieval," *International Journal of Man-Machine Studies,* vol. 18, pp. 407-439, 1983.

[14] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM,* vol. 18, pp. 613-620, 1975.

[15] J. Han and M. Kamber, *Data Mining - Concepts and Techniques*: Elsevier, 2006.

[16] W. Pedrycz and P. Rai, "Collaborative Clustering with the use of Fuzzy C-means and its Quantification," *Journal of Fuzzy Sets and Systems,* pp. 1-29, 2008.

[17] M. Ben Jlaiel, S. Kanoun, A. M. Alimi, and R. Mullot, "Three decision levels strategy for Arabic and Latin texts differentiation in printed and handwritten natures," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2007, pp. 1103-1107.