Using A Hidden Markov Model to Transcribe Handwritten Bushman Texts

Kyle Williams
Department of Computer Science
University of Cape Town
Private Bag X3, Rondebosch, 7701
South Africa
kwilliams@cs.uct.ac.za

Hussein Suleman
Department of Computer Science
University of Cape Town
Private Bag X3, Rondebosch, 7701
South Africa
hussein@cs.uct.ac.za

ABSTRACT

The Bushman texts in the Bleek and Lloyd Collection contain complex diacritics that make automatic transcription difficult. Transcriptions of these texts would allow for enhanced digital library services to be created for interacting with the collection. In this study, an investigation into automatic transcription of the Bushman texts was performed using the popular method of using a Hidden Markov Model for text line recognition. The results show that while this technique may be well suited to well-constrained and understood scripts, its application to more complex scripts introduces a number of difficulties that need to be overcome.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; I.7.5 [**Document and Text Processing**]: Document Capture—*Optical Character Recognition (OCR)*

General Terms

Algorithms, Experimentation, Performance

Keywords

OCR, handwriting recognition, Hidden Markov Model, digital libraries

1. INTRODUCTION

The Bleek and Lloyd Collection [5] contains notebooks, dictionaries and artwork that detail the language and culture of the Bushman people who lived in Southern Africa in the late 19th century. Transcriptions of these texts can be used to create enhanced digital library services for interacting with the collection. However, manual transcription is a costly and time-consuming task and many of the tools that have been developed for automatic transcription are not well suited to complex historical scripts.

In the case of the Bleek and Lloyd Collection, the script is complex due to the diacritics that appear in the text (see Figure 1) and that can appear above, below and both above and below characters and also can span multiple characters. Thus far about 137 different combinations of single and stacked diacritics have been discovered. These diacritics complicate the segmentation process and make training and recognition difficult due to the potentially large number of unique character/diacritic combinations and the similarities between characters that only differ due to diacritics.

Figure 1: Bushman diacritics

A popular approach to handwriting recognition is text line recognition using a Hidden Markov Model (HMM), as it allows for the complex segmentation of words and characters to be avoided. This paper describes an investigation of the application of text line recognition using an HMM for Bushman texts and discusses some of the problems that have been encountered along the way.

2. RELATED WORK

Indermuhle et al [3] built an HMM-based text line recognition system that compared a writer-specific HMM and an HMM based on general data, which was then adapted for a specific writer. Vinciarelli et al [6] built a text-line recognition system in which they did not differentiate between upper and lower case characters and considered symbols to be noise. Su et al [4] used an HMM-based text-line recognition system for recognising lines of Chinese text. Chinese script, like the Bushman script, is more complex than Latin-based scripts and Su et al trained a total of 2075 character models.

3. DESIGN

The system included a preprocessing step in which the pages in the Bleek and Lloyd Collection were binarised and the lines of the text in the pages segmented using the horizontal projection profile, which was based on the number of foreground-background transitions [1].

Since the Bushman script cannot be represented using standard Unicode, custom macros for the LaTeX TIPA package were created that allowed for the text to be represented. A custom AJAX tool, that allowed the user to capture the Latin characters of the Bushman text, add the diacritics and then preview the LaTeX markup, was built to allow for the capturing of the Bushman text that was used for training and recognition.

A recognition system was built that was based on a left-to-right continuous HMM with 12 emitting states and 16 Guassian components. Models were trained for each of the

JCDL'11, June 13–17, 2011, Ottawa, Ontario, Canada.
ACM 978-1-4503-0744-4/11/06.

character/diacritic combinations and for the characters ignoring diacritics and then combined to create word level models. Features were extracted using a sliding window approach where 9 geometrical features [1] and the first 3 coefficients (excluding the zero-frequency component) of the Discrete Cosine Transform [2] were extracted for every window. The system was tested using each of these feature sets independently as well as in combinations. No statistical language model exists for the Bushman languages and, as such, statistical information was not integrated into the recognition system.

4. EVALUATION

Evaluation of the system was performed using 1 of the 157 notebooks in the collection. The 74 pages in the notebook were automatically thresholded and then segmented into 995 separate text lines. Of these 995 lines, 297 were excluded from the training and testing phases due to noise, 111 were excluded due to bad segmentation and one sample was excluded because it could not be fully represented using the LaTeX representation used in this study. The system was evaluated using 10-fold cross validation using the remaining 698 samples, which were randomly allocated to 10 folds. The results reported are the averages of the results of the 10 folds.

The measures used for evaluating the system were the recognition rate and the recognition accuracy, where recognition rate is given by Equation 1 and recognition accuracy is given by Equation 2.

$$\text{Recognition Rate} = \frac{N - S - D}{N} \tag{1}$$

$$\text{Recognition Accuracy} = \frac{N - S - D - I}{N}, \tag{2}$$

where N is the length of the transcription, S is the number of substitutions, D is the number of deletions and I is the number of insertions needed to make the output of the recognition process match the correct transcription perfectly.

Three experiments were conducted as part of the evaluation. In the first experiment, HMMs were trained for each character/diacritic combination and character/diacritic combinations were recognised. In the second experiment HMMs were trained for each character/diacritic combination and diacritics were ignored during recognition, ie. if the base characters were correctly recognised then it was marked as correct regardless of any diacritics. In the third experiment, HMMs were trained using transcriptions that ignored diacritics and recognition was also performed ignoring diacritics.

4.1 Results

The recognition rate and accuracy for the character/diacritic combinations in Experiment 1 were the lowest, averaging at 56.34% and 43.84% respectively. It is speculated that the main reason for this is the large number of combinations of characters and diacritics that are possible and the lack of training samples for many of the models. The results from Experiment 2 showed that by ignoring the diacritics in the recognition process, the recognition rate and accuracy increased by about 10%, indicating that for about 10% of the character/diacritic combinations, the base characters are correctly recognised but the diacritics are not. Lastly, the recognition rate and accuracy for Experiment 3 were approximately 59.63% and 49.69% respectively, thereby demonstrating that there is a relatively high level of variation within models when diacritics are ignored, indicating that it is not a viable approach for use with a multilayer classifier.

5. DISCUSSION & CONCLUSION

In this study, a common approach to handwriting recognition was applied to the Bushman texts that appear in the Bleek and Lloyd Collection. These texts contain complex diacritics that make the recognition process difficult due to the number of character/diacritic combinations that exist. The effect of this complex representation is that, for the majority of cases, there are fewer than 3 training samples. It was shown that an improvement can be achieved when diacritics were ignored. However, the diacritics play an important role in the Bushman languages and as such cannot be ignored. The findings suggest that the common approach of text line recognition might not be well suited to the Bushman languages and motivates the need to investigate other techniques for recognition. These findings are not limited to the Bushman languages, but also apply to other complex scripts that are difficult to represent, and have meanings that are encoded in non-standard representations.

6. ACKNOWLEDGEMENTS

This research is supported by the University of Cape Town, the Telkom/NSN/Telesciences/THRIP Center of Excellence and the National Research Foundation.

7. REFERENCES

[1] On the influence of vocabulary size and language models in unconstrained handwritten text recognition. In *ICDAR '01: Proceedings of the Sixth International Conference on Document Analysis and Recognition*, page 260, Washington, DC, USA, 2001. IEEE Computer Society.

[2] N. Ahmed, T. Natarajan, and K.R. Rao. Discrete cosine transfom. *Computers, IEEE Transactions on*, C-23(1):90 – 93, 1974.

[3] Emanuel Indermühle, Marcus Liwicki, and Horst Bunke. Recognition of handwritten historical documents: HMM-Adaptation vs. writer specific training. In *Proceedings of The 11th International Conference on Frontiers in Handwriting Recognition*, 2008.

[4] Tong-Hua Su, Tian-Wen Zhang, De-Jun Guan, and Hu-Jie Huang. Off-line recognition of realistic chinese handwriting using segmentation-free strategy. *Pattern Recogn.*, 42(1):167–182, 2009.

[5] Hussein Suleman. Digital libraries without databases: The bleek and lloyd collection. In *Research and Advanced Technology for Digital Libraries*, pages 392–403. 2007.

[6] Alessandro Vinciarelli, Samy Bengio, and Horst Bunke. Offline recognition of unconstrained handwritten texts using hmms and statistical language models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(6):709–720, 2004.