Learning to Read Bushman

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 notebooks in the Bleek and Lloyd collection contain handwritten stories that metaphorically encode the Bushman culture and are useful to researchers and scholars trying to understand Bushman language and culture. These notebooks, however, only exist as scanned images and therefore the stories they contain cannot be searched, indexed or compared. This research seeks to investigate how accurately the Bushman stories can be automatically converted from images to text, in a process known as transcription, and also to explore the various techniques for doing this. The expected contribution is a measurement of how accurately transcription can be automatically performed as well as a comparison of different techniques for doing this.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; I.2.7 [**Artificial Intelligence**]: Natural Language Processing; I.7.5 [**Document and Text Processing**]: Document Capture

General Terms

Algorithms, Experimentation, Performance

Keywords

OCR, machine learning, handwritten manuscripts, cultural heritage preservation, digital libraries

1. INTRODUCTION

The notebooks in the Bleek and Lloyd Collection [4] are available to researchers and scholars as digital scans. While these digital scans are useful for learning about Bushman language and culture, they do not allow for the automatic indexing, searching and comparison of the text that appears in them. Transcription will allow these and other operations to be performed. However, manual transcription is a tedious and impractical task due to the size of the collection. This research deals with the automation of this task.

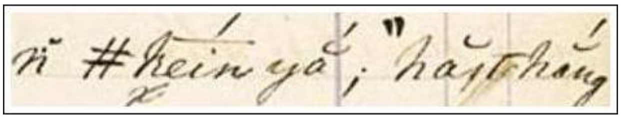

Figure 1: Diacritics above, below and spanning multiple characters

2. OBJECTIVES

Handwriting transcription is a well studied problem [1]. However, for the Bushman languages, it is a difficult task due to the diacritics that appear above and below characters and that span multiple characters (Figure 1). The lack of a language model also contributes to the difficulty of the task. The diacritics complicate the recognition and segmentation processes, while the lack of a language model does not allow for the incorporation of knowledge about the language in order to improve results. The objectives of this research are to determine how accurately handwritten Bushman texts can be automatically transcribed and which of the techniques used in the literature are most effective for doing this. In this sense, the objectives include determining which combinations of segmentation techniques, features and machine learning algorithms perform best, as well as determining the extent to which a language model can be built and used to improve results.

3. METHODOLOGY

3.1 Research Questions

An experimental methodology will be used in this research in order to answer the following main research question and sub-research questions.

- How accurately can the automatic transcription of handwritten Bushman texts be performed?
  - Which of a selection of Hidden Markov Models, Neural Networks and Support Vector Machines, when used in conjunction with various feature sets, performs best when automatically transcribing handwritten Bushman texts?
  - Which segmentation techniques are effective for the machine learning algorithms used in this research?
  - To what extent can an n-gram language model improve accuracy when automatically transcribing handwritten Bushman texts?

The main research question seeks to investigate the general overall accuracy with which the automatic transcription of handwritten Bushman texts can be performed,

SAICSIT '10, October 11-13, Bela Bela, South Africa

while the sub-research questions investigate the various ways to go about doing it.

3.2 Development and Evaluation

To investigate which machine learning algorithms and feature sets perform best, Hidden Markov Models, Neural Networks and Support Vector Machines will be built making use of various feature sets. These machine learning algorithms and feature sets will then be evaluated by directly comparing transcription accuracy, word error rates and other statistical measures for differences among combinations.

Various segmentation techniques will be employed in order to to investigate which segmentation techniques are the most effective for the different machine learning algorithms used in this study. These segmentation techniques will be evaluated based on the effect that they have on transcription accuracy, word error rates and other statistical measures for differences among combinations.

An n-gram language model will be built and incorporated into the system. Doing this will allow for an evaluation to be performed to determine the extent to which an n-gram language model can be used to improve accuracy when automatically transcribing handwritten Bushman texts. This improvement can be directly measured by comparing transcription accuracy and word error rates with and without the n-gram language model.

The main research question concerning the accuracy with which handwritten Bushman texts can be automatically transcribed can be determined based on the evaluation above. Furthermore, additional end-to-end evaluation will be conducted to gain insight into the tradeoff between accuracy and performance and other important considerations in the transcription process.

Evaluation will be performed using K-fold cross-validation on a representative set of the collection. The data in the representative set will be decided on in discussion with experts from the Michaelis School of Fine Art.

3.3 Pilot Study

A pilot study was conducted to investigate the feasibility of the automatic transcription of Bushman characters [5]. In the pilot study, the automatic transcription of a limited set of Bushman characters was conducted for a neatly rewritten version of a Bushman story. In the pilot study an accuracy rate of approximately 80% was achieved, when using a training set created by two authors, under-sampled bitmaps [3] as features and a support vector machine [2] for classification and recognition. The pilot study provided useful insight into the problem of automatically transcribing handwritten Bushman texts, however, it over-simplified many of the problems associated with automatically transcribing handwritten Bushman texts.

4. EXPECTED OUTCOMES

The main expected outcome of this research is a measurement of the accuracy with which automatic transcription of Bushman texts can be performed. Another expected outcome is a comparison of a selection of Hidden Markov Models, Neural Networks and Support Vector Machines used to classify and recognise Bushman languages and insight into which feature sets are best suited to handwritten characters with diacritics. Furthermore, insight will be gained into effective segmentation approaches for the different machine learning algorithms considered.

A software system will be produced that is capable of transcribing handwritten Bushman texts using a variety of features and the three machine learning algorithms previously mentioned. Additionally, a transcription of some of the Bleek and Lloyd notebooks is expected, with a high level of accuracy.

Success will largely be judged based on the research providing a strong and convincing comparison of the various feature sets and machine learning algorithms considered.

5. PRACTICAL IMPLICATIONS

It is expected that this research could have a number of practical implications, such as being used in text-to-speech applications, reprinting of the stories that appear in the Bleek and Lloyd Collection and allowing for automatic translation of the texts using the Bleek and Lloyd dictionaries [6]. Furthermore, it is expected it will allow researchers and scholar to gain insight into Bushman culture and enhance their understanding of some of the earliest inhabitants of the Earth. Lastly, it is expected that this research will have practical implications for other researchers in the field, especially those working with texts that contain complex diacritics.

6. CONCLUSIONS

The aims of this research are to investigate the best techniques for automatically transcribing handwritten Bushman texts. The various techniques used in the literature will be compared to determine which are most effective for Bushman texts and which combinations of techniques work best together. It is expected that the findings from this research will not only be applicable to the Bushman texts which appear in the Bleek and Lloyd collection, but also other texts which contain complex diacritics.

7. ACKNOWLEDGMENTS

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

8. REFERENCES

[1] H. Bunke. Recognition of cursive roman handwriting: past, present and future. In *Document Analysis and Recognition*, pages 448 – 459 vol.1, 2003.

[2] C. Cortes and V. Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[3] A. Oliveira, C. Mello, E. Silva Jr, and V. Alves. Optical digit recognition for images of handwritten historical documents. In *Neural Networks, 2006. SBRN '06. Ninth Brazilian Symposium on*, pages 166 –171, oct. 2006.

[4] H. Suleman. Digital libraries without databases: The bleek and lloyd collection. In *11th European Conference on Research and Advanced Technology for Digital Libraries*, pages 392–403, March 2007.

[5] K. Williams. Feasibility of automatic transcription of neatly rewritten bushman texts. Technical report, Department of Computer Science, University of Cape Town, 2010. Technical Report CS10-06-00.

[6] K. Williams, S. Manilal, L. Molwantoa, and H. Suleman. A visual dictionary for an extinct language. In G. Chowdhury, C. Koo, and J. Hunter, editors, *The Role of Digital Libraries in a Time of Global Change*, volume 6102 of *Lecture Notes in Computer Science*, pages 1–4. Springer, 2010.