Feasibility of Automatic Transcription of Neatly Rewritten Bushman Texts

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

Technical Report CS10-006-00

April 2010

Introduction

The purpose of this study is to investigate the feasibility of the research to be conducted for a MSc. The study is concerned with the automatic transcription of part of a handwritten |xam story, which contains a limited set of characters from the |xam Bushman language. The transcription is performed using a trained SVM [1] model to classify the characters. The text to be transcribed is a neatly rewritten version of the first page of *A Story of the Girl who made the Milky Way*, which appears in one of Lucy Lloyd's |xam notebooks. Two authors participated in this study with the purpose of evaluating the ability to transcribe the handwriting of multiple authors.

Data Description

A Story of the Girl who made the Milky Way, which appears in one of Lucy Lloyd |xam notebooks was used in this study. A scan of the story from the notebook is shown in Figure 1. There are 39 different characters (classes) in this story which are shown in Appendix A.

There are two types of data in this study: training data and testing data. Two authors participated in this study and created both the training data and the testing data. Author 1 provided 10 examples of each character in the story as training data and author 2 provided 3 examples of each character in the story as training data. Each author then neatly wrote out the story which was then used as testing data for transcription. The training data for one character is shown in Figure 2 and the neatly handwritten story by each author is shown in Figure 3. An example of the story written in English, |xam and classified into character classes is shown in Appendix B.

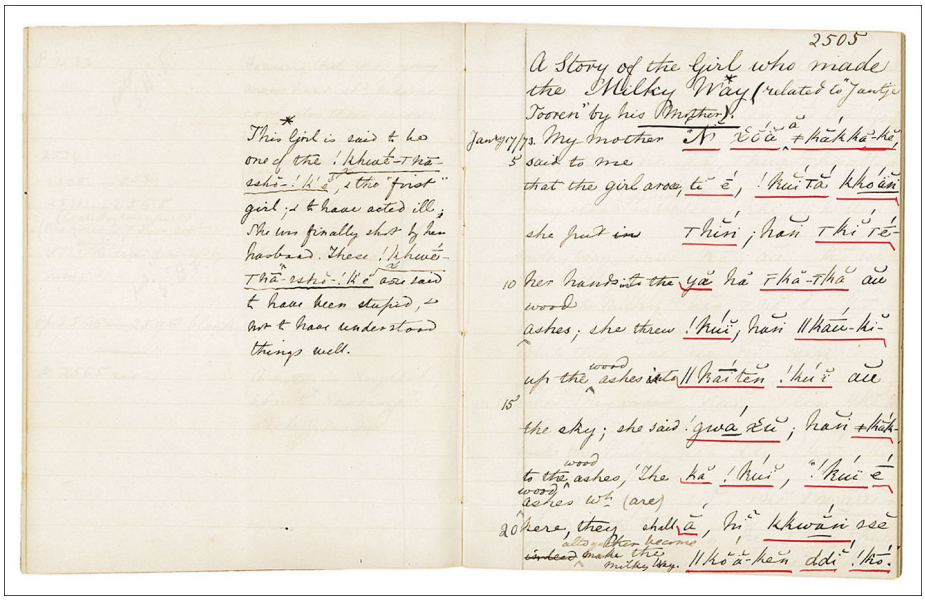


Figure 1: The original notebook image on which the story appears.

(a) Training data for author 1

(b) Training data for author 2

Figure 2: Training data for study

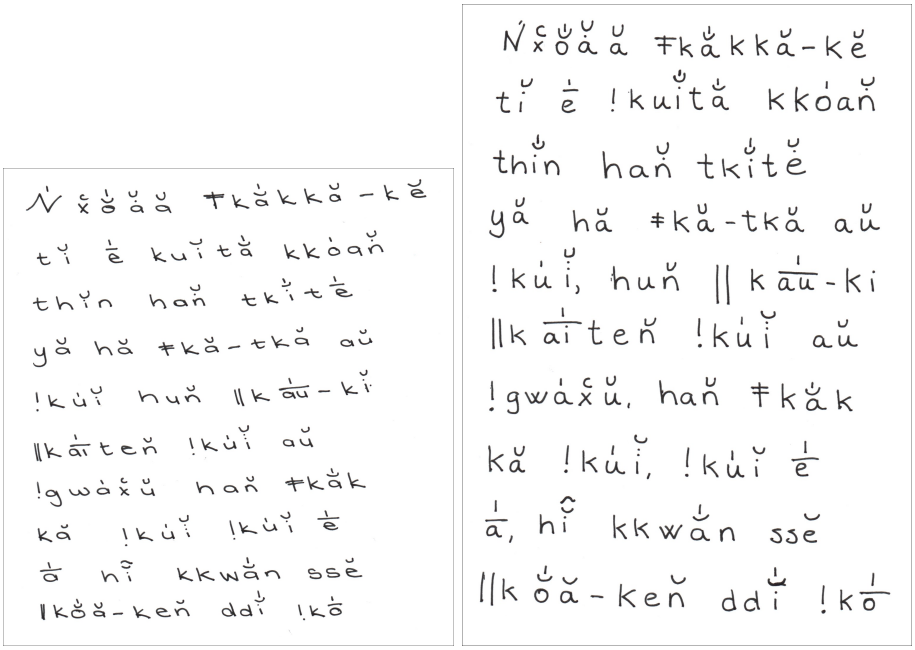
Implementation

There are 4 steps involved in the implementation of the system: segmentation; feature extraction; model training; and transcription. These steps will each be discussed in more detail here.

Segmentation

Segmentation involves separating a page of text into lines, lines into words and words into characters. Each class of training data was provided as a segmented line for which each character had to be segmented. This segmentation was done using connected component analysis [3]. In cases where segmentation was wrong, errors were manually corrected.

For segmenting the neatly written pages containing the story, line segmentation was performed, then, for each line, word segmentation was performed and, for each word, character segmentation was performed. This segmentation was performed using connected component analysis [3]. In cases where segmentation errors occurred, the full story pages were manually edited until they could be perfectly segmented automatically. This editing usually involved increasing or decreasing the spaces between words and characters.



(a) Testing data for author 1

(b) Testing data for author 2

Figure 3: testing data for study

Feature Extraction

Each character which was segmented was normalised by resizing it to 32x32 pixels and thresholding the image. For each normalised image, a 4x4 sliding window counted the number of black pixels in the window, thereby creating an 8x8 feature matrix [4]. This feature matrix was then normalised over the range [0-1].

Model Training

An SVM was trained using the features for each training sample and an RBF kernel with parameters C and γ. The training data for the SVM was scaled and the values for the parameters C and γ were found using 5-fold cross validation. A total of 3 models were trained:

1. Model A1: Author 1 - 388 examples.
2. Model A2: Author 2 - 117 samples.
3. Model A1_2: Authors 1 & 2 - 505 samples.

Transcription

Transcription involves taking an image as input, automatically segmenting the lines, words and characters, extracting features from each character, scaling the data and then using the SVM model to transcribe the image.

1 Evaluation

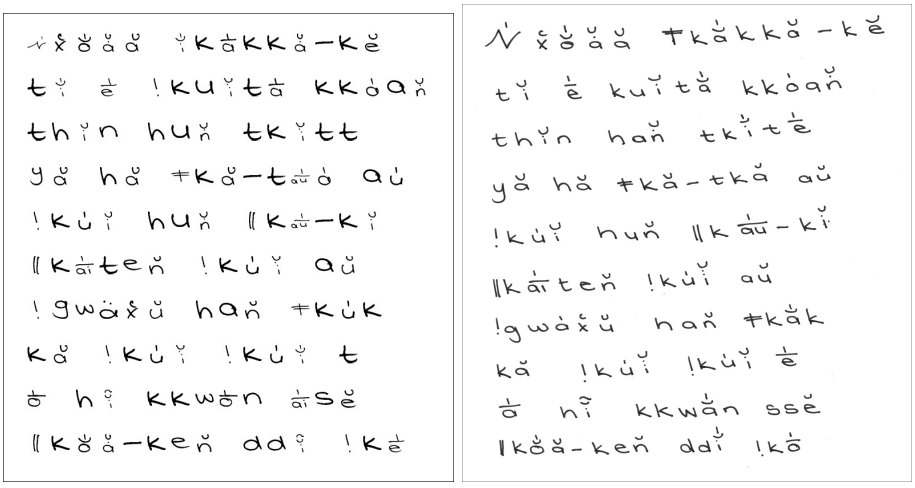
A total of 6 cases were evaluated in this study:

- Story by Author 1 and Model A1.
- Story by Author 2 and Model A1.
- Story by Author 1 and Model A2.
- Story by Author 2 and Model A2.
- Story by Author 1 and Model A1_2.
- Story by Author 2 and Model A1_2.

Table 1 shows the results of the evaluation and Figure 4 shows the transcription which achieved the highest accuracy alongside the correct transcription.

Table 1: Accuracy of Transcription (%)

	Model A1	Model A2	Model A1_2
Author 1	77.1654%	28.3465%	79.5276%
Author 2	36.2205%	62.9921%	71.6535%



(a) Output of transcription for Author 1 with Model A1_2

(b) Correct transcription

Figure 4: Comparison of most accurate transcription with correct transcription.

The results show that using an author's training set to transcribe that author's handwriting can be done with a satisfactory level of accuracy - approximate 78% and 63% for authors 1 and 2 respectively. Similarly, the results show that using one author's training set to transcribe another author's handwriting results in poor levels of accuracy. Transcribing the handwriting of author 2 using author 1's training set only led to approximately 36% accuracy, however, by augmenting the relatively large training set of author 1 with the relatively small

training set of author 2, the accuracy increased to approximately 72%. This finding is in line with the findings in [2] in which Hidden Markov Models were used to transcribe handwritten historical documents and where it was shown that augmenting a general training set with author specific training samples improved accuracy. The highest accuracy, approximately 80%, was achieved using the augmented training set of both authors, a possible reason for this being that the augmented training set accounts for more variability due to the different writing styles of the authors.

Conclusions

This study set out to investigate the feasibility of the automatic transcription of neatly rewritten |xam characters. An SVM was trained using data from two authors and then the SVM was used to transcribe an image of a story. The highest accuracy achieved was approximately 80% and this occurred when the training data from both authors was used to train the SVM. The findings suggest that using multiple authors could have a positive effect on transcription, since multiple authors increase the variability of the training data - this, however, is not conclusive and needs to be investigated further. Diacritics in the text have the potential to cause problems both at the segmentation level and at the training level. Overall, this study suggests that the transcription of handwritten Bushman texts is possible, though special attention needs to be paid to issues arising from diacritics as well as dealing with multiple authors.

References

[1] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[2] Horst Bunke. Emanuel Indermhle, Marcus Liwicki. Recognition of handwritten historical documents: Hmm-adaptation vs. writer specific training. In *Eleventh international Conference on Fronteirs in Handwriting Recognition*, 2008.

[3] T. Yung Kong and Azriel Rosenfeld, editors. *Topological Algorithms for Digital Image Processing.* Elsevier Science Inc., New York, NY, USA, 1996.

[4] A.L.I. Oliveira, C.A.B. Mello, E.R. Silva Jr, and V.M.O.Alves 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.

A Characters in Story

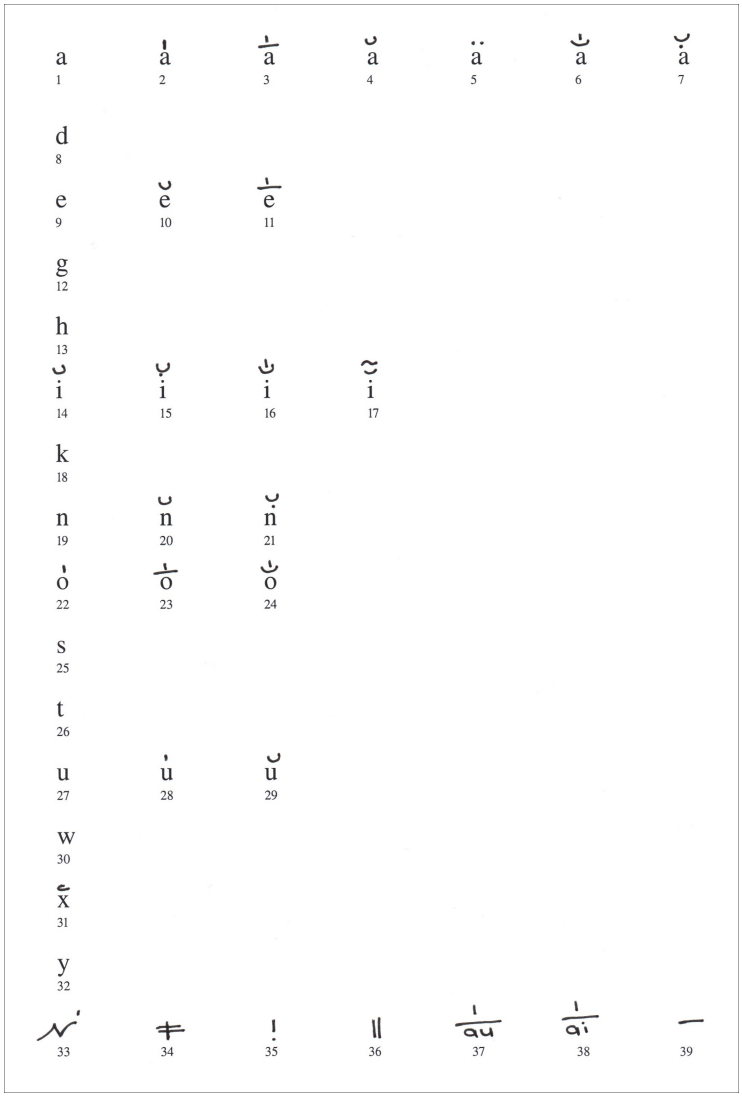
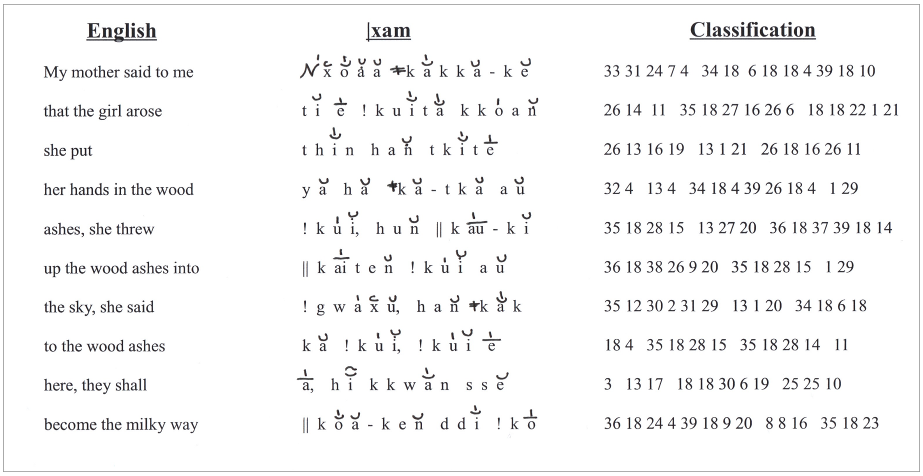


Figure 5: Characters that appear in the text and that are used in the study. There is a total of 39 character classes.

B Story in English, |xam and Classified Characters



English	|xam	Classification
My mother said to me	Nxoaa ǂkakka-ke	33 31 24 7 4 34 18 6 18 18 4 39 18 10
that the girl arose	ti e !kuita kkoan	26 14 11 35 18 27 16 26 6 18 18 22 1 21
she put	thin han tkite	26 13 16 19 13 1 21 26 18 16 26 11
her hands in the wood	ya ha ǂka-tka au	32 4 13 4 34 18 4 39 26 18 4 1 29
ashes, she threw	!kui, hun ||kau-ki	35 18 28 15 13 27 20 36 18 37 39 18 14
up the wood ashes into	||kaiten !kui au	36 18 38 26 9 20 35 18 28 15 1 29
the sky, she said	!gwaxu, han ǂkak	35 12 30 2 31 29 13 1 20 34 18 6 18
to the wood ashes	ka !kui, !kui e	18 4 35 18 28 15 35 18 28 14 11
here, they shall	a, hi kkwan sse	3 13 17 18 18 30 6 19 25 25 10
become the milky way	||koa-ken ddi !ko	36 18 24 4 39 18 9 20 8 8 16 35 18 23

Figure 6: English, |xam, and transcribed/classified versions of the story.