A System for Historical Documents Transcription based on Hierarchical Classification and Dictionary Matching

Camelia Lemnaru, Andreea Sin-Neamțiu, Mihai-Andrei Vereș and Rodica Potolea
Technical University of Cluj-Napoca, 26-28 Barițiu Street, Cluj-Napoca, Romania

Keywords: Handwriting Recognition, Historical Document, Hierarchical Classifier, Dictionary Analysis, Kurrent Schrift.

Abstract: Information contained in historical sources is highly important for the research of historians; yet, extracting it manually from documents written in difficult scripts is often an expensive and time-consuming process. This paper proposes a modular system for transcribing documents written in a challenging script (German *Kurrent Schrift*). The solution comprises of three main stages: Document Processing, Word Processing and Word Selector, chained together in a linear pipeline. The system is currently under development, with several modules in each stage already implemented and evaluated. The main focus so far has been on the character recognition module, where a hierarchical classifier is proposed. Preliminary evaluations on the character recognition module has yielded ~ 82% overall character recognition rate, and a series of groups of confusable characters, for which an additional identification model is currently investigated. Also, word composition based on a dictionary matching approach using the Levenshtein distance is presented.

1 INTRODUCTION

The process of transcribing historical documents requires the expertise of paleographers, due to the large variety of languages and scripts, as well as the (low) quality of the manuscripts (Minert, 2001). A paleographer, specialized in specific forms of writing, performs the transcription by hand, a tedious process that takes a considerable amount of time and effort. This suggests the need for an automated process capable of performing the transcription with minimal user intervention, reducing the costs of a transcription.

The problem of transcribing a historical document (given the historical period, language and script) can be placed in the area of pattern recognition, namely the problem of handwriting recognition. Two handwriting recognition techniques are most commonly considered (Fischer, 2010): on-line recognition (performed on-the-fly with the aid of an active surface and a pen) and off-line recognition (which extracts the written text from an input raster image). As historical documents are converted into digital format via individual scans, the transcription process is an off-line handwriting recognition task.

Handwriting recognition has been a research subject for almost a decade, nowadays being considered an important subject for both pattern recognition and data mining fields. Current solutions focus on either dynamic user input mechanics (commonly integrated in modern devices with touchscreen capabilities) or forensic-related identifications. The field of historical handwritten document transcriptions is currently a valuable research problem.

The system proposed in (Fischer, 2010) considers *Medieval Documents*, in which text isolation is difficult due to poor paper quality. Medieval handwriting however displays reduced irregularities, with small in-class variations and uniform text localization. The solution proposed in (Juan, 2010) focuses on an interactive transcription environment. The approach assists the paleographer, by providing advanced user interaction capabilities and preserving the topology of the original document thorough the transcription. It does not aim to produce an automated process and is not suitable for historians unfamiliar with the work of transcription.

Lemnaru C., Sin-Neamţiu A., VereÈŹ M. and Potolea R..
A System for Historical Documents Transcription based on Hierarchical Classification and Dictionary Matching.
DOI: 10.5220/0004143003530357
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2012), pages 353-357
ISBN: 978-989-8565-29-7

2 PROPOSED SYSTEM

We designed an automated system capable of successfully transcribing documents written in German *Kurrent Schrift*. We claim that the system can be easily adapted to support simpler scripts (such as Latin, or Greek). The solution assumes that complex restoration filters which enable proper text isolation are not necessary (i.e. the image quality of the documents is fair). However, light imperfections (such as faint paper folding, material aging and isolated ink droplets) are managed by the system (example of document presented in Figure 1).

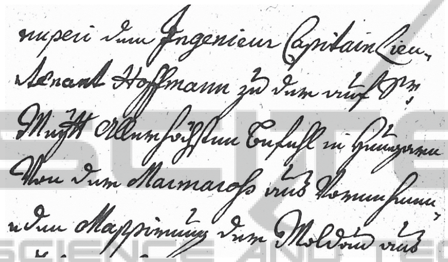

Figure 1: Excerpt from a Kurrent Schrift historical document.

2.1 Conceptual Architecture

We have identified three major stages – Document Processing, Word Processing and Word Selection, connected in a linear pipeline (as seen in Figure 2).

First, the document image (i.e. the scanned version of the historical handwritten document) is processed by the ***Document Processor***. After separating the text from the background through a two-step procedure and removing malignant noisy areas, the document is de-skewed to improve the correctness of subsequent processing steps. It is then successively partitioned into lines of text and individual words.

The ***Word Processor***, the core component of the system, finalizes the preprocessing of the input word images, by performing slant correction and character splitting. The shape of the binary character objects is then captured using a skeletonization filter, and important features that discriminate the characters are extracted. A classifier identifies each character and word variants are constructed.

The words are validated by the ***Word Selector*** using a local dictionary database and a Knowledge Base, generating transcription variants with attached probability. Inappropriate matches are pruned and the words reordered such as to generate the final transcription, the output of the system.

2.2 Component Description

The *Document Processor* extracts the text from the background using a binary conversion of the image in a two-step process: greyscaling followed by binarization. Global thresholding (Otsu, 1979) offers the best performance – computational complexity ratio. Noise reduction is ensured by a blobs-based labelling technique, removing objects of having the area smaller than a threshold value (dependent on the image size).

Due to the fact that the human writer most often fails to write text on perfectly horizontal lines, text is written at a certain angle. Individual characters are therefore distorted. This problem is commonly known as document skew, which we attempt to minimize through a Projection-Based correction (Zeeuw, 2006), considering multiple skew angles. The actual correction is performed by a vertical shearing in the opposite direction of the skew (Sun, 1997).

Line splitting is performed by applying a Gaussian smoothing filter to detect horizontal projection areas of low density. Split points are identified as local minima inside a rectangular centred window and are separated from neighbouring split points by a peak. Analogously, lines of text are separated into words based on the individual vertical projections.

Due to the possibility of having irregular word orientation inside the line of text, the *Word Processor* performs slant-correction, similar to skew-correction through horizontal shearing. Words are then vertically cut into individual characters (Zeeuw, 2006).

The extraction of the shape of the binary characters is done by thinning (using a skeletonization filter). K3M thinning (Saeed, 2010) is employed, which generates a pixel-width, connected skeleton. Pruning of spurious branches ensures a stable skeleton structure (Bai, 2005), unaffected by small shape variations.

Significant numerical features are extracted from the resulting shape in order to discriminate the characters, considering both strong inter-class variation and weak correlations (avoidance of redundancies). The following mix of features is considered (Vamvakas, 2007): projections, profiles, transitions and zone densities. Because histogram-based features are dependent on the character image resolution (width, height), we propose histogram-compression based on the Discrete Cosine Transform. This approach captures the shape of the histogram.

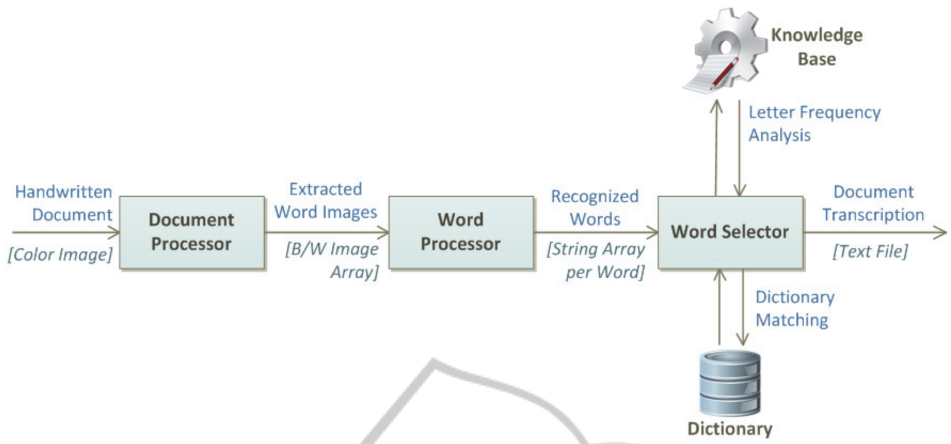

Figure 2: Conceptual architecture.

The resulting sequence is then normalized in the [0,1] interval, and only the first few coefficients are considered. This ensures both a fixed-dimension feature vector and noise reduction (generally located in the higher part of the signal's spectrum). Parameter tuning is possible by varying the number of considered harmonics (coefficients), the optimal truncation index being determined through experiments.

We propose a hierarchical two-level classifier for character identification: at the first level, easily distinguishable characters are identified, while at the second level confusable groups of characters are considered. The second level classifier consists of an array of classifiers, each specialized on a certain confusable group.

The final result is computed as a weighted normalized sum, considering probabilities from both levels. The output of the character classification module is the Probability Density Function (PDF) of all classes. Due to possible misclassifications for highly similar characters, several highly ranked characters are returned which are combined into words. Those with large probabilities (above a threshold) are passed to the last module.

Because there are situations in which not even the human eye can identify handwritten characters in an isolated manner, word context is necessary to reduce ambiguity and increase system accuracy. Therefore, the *Word Selector* validates the word variants using a dedicated dictionary. The proposed approach searches for similar words using the Levenshtein Distance. We expect the most frequent case to appear will be that of partially/improperly recognized words with correctly identified length.

Finally, a score generator decides among the words by considering both the recognition probability and the Levenshtein Distance to the dictionary word, as in formula (1):

$$score(W) = \omega\, p(W) + (1-\omega) Max\left(0, \left(1 - \frac{Lev(W)}{Len(W)}\right)\right) \quad (1)$$

Lev(W) – Levenshtein Distance between initial word and dictionary solution

Len(W) – Length of initial word

The value of the weight factor ω controls the relative importance of the two factors – recognition rate and Levenshtein Distance and is determined empirically. Based on the final score, the best solution is placed into the transcription, with possible variants also available to the user.

3 EXPERIMENTS AND RESULTS

The majority of experiments performed so far have focused on the character identification tasks, in order to find the most accurate classifier for primary character identification. Tests on the *first level* of the classifier were performed using a balanced data set having 25 classes and 37 features. Stratified 10-fold cross-validation was considered, employing accuracy as performance metric. We started with representative classifiers which were expected to provide good results due to their robustness in multi-class problems. Parameter tuning improved the performance of two of them (Table 1).

The tests on the Multilayer-Perceptron focused on variations of the learning rate (0.2-0.4), momentum (0.1-0.5), validation threshold (15-25) and hidden layers structure. Lower learning rate as well as single hidden layer configurations yielded the best results. The best configuration features a learning rate of 0.2, one hidden layer of 62 cells and a momentum of 0.1.

The behaviour of the Support Vector Machine (SMO) was tested with respect to complexity variation (1.0 – 6.0) and various kernels (PUK, RBF, PolyKernel, Normalized Poly). The best results were obtained for PUK and RBF kernels, and a rather high complexity (5.5 and 6, respectively).

The most relevant parameter for the Random Forest classifier is the number of trees (10-60). Above the value of 35 trees, the classifier presented oscillatory accuracy, being suspicious of overfitting.

Naïve-Bayes improvements were attempted using Adaptive Boosting (10-60 iterations, 80-120 weight threshold, with or without resampling). The resulting classifiers exhibited no accuracy improvement.

The influence of the feature vector size (induced by the number of computed DCT coefficients) on the performance has also been evaluated for the best classifiers (Figure 3). The results indicate that the optimal number of DCT coefficients lies between 4 and 6.

Table 1: Classifier experiments summary.

Classifier	Initial Accuracy	Parameter Tuning
MLP	72.51%	75.29%
SMO	71.53%	79.60%
RF	67.69%	Overfitted behaviour
Naïve Bayes	66.06%	No improvement

The validity of the proposed set of features has been confirmed by applying two feature selection techniques (SVM and Gain Ratio attribute evaluator rankers from WEKA). All features were considered valid, with the Ranker unable to group the DCT coefficients coherently based on the primary features.

A series of preliminary experiments were performed on the Word Selector module, using a 687 words dictionary. Partially-recognized words were considered, for which the system yielded successful identifications, as exemplified in Table 2. Further experiments are required for a larger dictionary database and other incompleteness patterns.

4 CONCLUSIONS

This paper proposes a new system for historical document transcription, which employs three main modules connected in a continuous pipeline: Document Processor, Word Processor and Word Selector.

The system is currently under development. Several sub-modules have been implemented and evaluated. The main focus so far has been on the character identification and word composition tasks, but several image processing steps have also been considered, such as image binarization, document segmentation, and stable skeletonization.

Extensive evaluations conducted for the character identification task using various machine learning methods and parameter values have yielded a best identification rate of ~82%, and a set of confusable characters for which a second layer classifier is currently being developed. To extend the recognition process in the case of individual characters impossible to be identified in an isolated context, a word-level dictionary analysis is employed. Preliminary results indicate a good identification for partially identified words using a Levenshtein-based search.

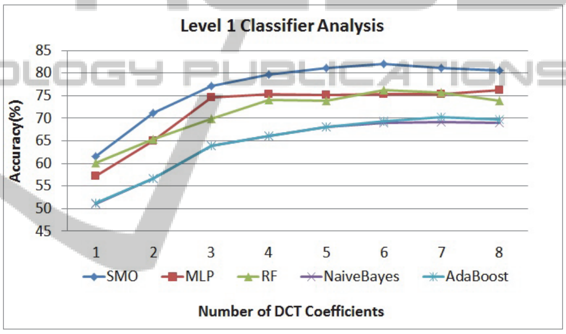

Figure 3: Impact of DCT coefficients on classifier accuracy.

Table 2: Dictionary matching example.

Word	Distance 2	Distance 3
a**h	auch	17 variants: nach, noch, sich, aber, etc
zuru**	zuruck	Zur

REFERENCES

Bai, X., Latecki, L. J., Liu, W., 2005. Skeleton Pruning by Countour Partitioning with Discrete Curve Evolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 3, March 2007

Fischer, A., Wüthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., Stolz, M., 2010. Automatic Transcription of Historical Medieval Documents. *DAS '10 Proc. of the 9th IARR*.

Juan, A., Romero, V., Sánchez, N. Serrano, J. A., Toselli, A. H., Vidal, E., 2010. Handwriting Text Recognition for Ancient Documents. *Workshop and Conference Proceedings 11-Workshop on Applications of Pattern Analysis*.

Minert, R. P., 2001. *Deciphering Handwriting in German Documents: Analyzing German, Latin and French in Vital Records Written in Germany*. GRT Publications.

Otsu, N.,1979. A threshold selection method from grey level histogram, *IEEE Transactions on Systems, Man, and Cybernetics*, vol SMC-9, No 1.

Saeed, K., Tabedzki, M., Rybnik, M., Adamski, M., 2010. K3M: A Universal Algorithm For Image Skeletonization And A Review Of Thinning Techniques. *Applied Mathematics and Computer Science*, Vol. 20, Nr2, p. 317-335.

Sun, C., Si, D., 1997. Skew and Slant Correction for Document Images Using Gradient Direction ICDAR 1997- *4th International Conference Document Analysis and Recognition.*

Vamvakas, G., 2007. *Optical Handwritten Character Recognition.* National Center for Scientific Research "Demokritos" Athens, Greece

Frank de Zeeuw, 2006. *Slant Correction using Histograms.* Bachelor's Thesis in Artificial Intelligence.