Novel Algorithmic Approach to Deciphering Rovash Inscriptions

Loránd Lehel Tóth
Budapest University of Technology and Economics, Hungary

Raymond Pardede
Budapest University of Technology and Economics, Hungary

Gábor Hosszú
Budapest University of Technology and Economics, Hungary

INTRODUCTION

In the Middle Ages, the Szekely-Hungarian Rovash script was carved in stone, paper, wood, and the walls of buildings, which were exposed to the weather for centuries. As a result, the paleographers usually encounter the problem of interpreting the carved text because of the incomplete graphemes or grapheme errors. The Rovash (pronounced "rove-ash," other spelling: Rovas) is a script family, which was used by nations in the Carpathian Basin and in the Eurasian Steppe. One member of this script family is the Szekely-Hungarian Rovash, which was used and gradually developed by the Szekelys in Szekelyland (present day Romania) (Hosszú, 2013).

The article introduces the methods used in the computerized paleography for identifying unknown inscription. The script identification differs from the Optical Character Recognition (OCR), since in the OCR, the normalized glyph of each grapheme belonging to a certain script are known (Szűcs, 2009). Therefore, the task of the OCR is to convert the signs in a certain inscription into the well-known normalized glyphs of an alphabet. In other words, the OCR focuses on the automatic grapheme extraction from a certain inscription. Oppositely, in the computerized paleography and more specifically, in the script identification the right interpretation of the signs in an inscription is the main problem to be solved. In several cases, the shape of the signs in an inscription can be copied easily to a sheet of paper; however, the inscription remains undeciphered. Its reasons can be the following: (1) the script used for making the inscription is unknown, (2) even if the script is known, but the normalized glyphs of the script is unknown (which can be specific for a certain age and a certain area), or (3) the language of the inscription is uncovered. Therefore, the script identification focuses on these three problems. Naturally, in several practical cases, the OCR and the script identification can be overlapped.

After shortly presenting the script identification methods, the article describes a novel script identification algorithm and its implementation called SID software. This algorithm is general purposed; however, it was exclusively applied for the Szekely-Hungarian Rovash script (Hosszú, 2011). The results obtained from the SID are also presented. Finally, the conclusions summarize the new method and the experimental results.

BACKGROUND

The Methods of the Computerized Paleography

From the 15th century, ink and paper became dominant writing technologies; however, due to the mutilation some parts of the scripts written on paper are fractional and difficult to read. Reading and understanding of the writings made in ancient times is difficult to many of the researchers. The reason for this apart the loss of the writing support material (wood, stone, brick, paper, etc.) is that the forms of the letters used in these writings have changed over the time, as well as the fact that different ancient writings were wrote in

DOI: 10.4018/978-1-4666-5888-2.ch711

different handwriting and by people with a different writing skills.

The international literature conducts extensive research in this area. Apart other things, valuable achievements were made in India among Kannada language speakers, in the algorithmic definition of the Kannada ancient writings used from 5th century AD (Kashyap et al., 2003). Their procedure attempts to identify the forms of the letters (glyph) in any text written in different historical periods with the help of the neural networks. To identify the digits that belong to Kannada writings in India the researchers used multi-level classifiers and fuzzy logics (Dinesh et al., 2009).

Another approach to decipher Egyptian hieroglyphics writing, the researchers have also developed a meaning identification algorithm in which the system receives an image of hieroglyphics as an input, and as an output it provides the translation of hieroglyphs (Jón, 2009). The noisy input is usually an obstacle in the printed or handwritten character recognition programs, which in addition to the output error often compromises the identification of detailed punctuation. To avoid this, the hieroglyph-recognition system uses such noise cancellation algorithm, which can remove hieroglyph non-related sets of pixels. The hieroglyph character recognition has been implemented by using the three-layer neural network. The system intelligence is carried by Feed-Forward Back-Propagation procedure applied by neural network, which uses a rapid learning algorithm and apart from the difficulties at initial stages of learning, it provides quite a quick positive results. The use of this method is also justified by the fact that the procedure already applied in other areas and the underlying mathematical background has a wide literature (Le Cun et al., 2009).

Similarly promising results have been achieved by another research group in process development of Old Persian texts acquisition. Just like the other systems, a method for the Persian handwriting recognition consists of the following steps: preprocessing, dismantling components, recognition, and control. The preprocessing phase includes an image processing (such as noise filtering, edge smoothing, filling and image transformation). Components dismantling aim is to recognize the symbols and their separation. Then, the detected symbols are assigned to class label and probability value. The control optional module approves the recognition variants and tries to reduce the error rate (Doermann & Jaeger, 2006). The algorithm takes into consideration structure and shape of the letters (Izadi et al., 2008).

One possible way of grapheme modeling is a multi-layer model, which consists of Topological, Visual Identity and Phonetic layers (Pardede et al., 2012). The researchers approach the meaning of identity of a hard to read archaic text in a three aspects.

The topological layer contains topological matrices, which describes some graphemes with topological components. A topological element of a grapheme can a closed loop, a vertical line, a horizontal line, an endpoint, etc., which can be visually recognized and identified. The meaning of the Visual Identity Layer is that it identifies the object visually, in terms of the human visual system. It provides a possible answer to the question: which unknown grapheme resembles a known one the most in a given script, taking into account the main topological properties.

Sound value correspondence to a grapheme is implemented in Phonetic Layers (Pardede et al., 2012). While studying ancient scripts it can be determined that different sound values can correspond with the same grapheme. The sound value correspondence varies depending on ages, geographic regions and authors. However, it can be determined that in one source the same sound value belongs to one grapheme.

Our novel algorithm is based on the specially selected topological properties of glyphs, and the difference between the symbols of the inscription and the graphemes of a certain alphabet is calculated.

The Features of the Szekely-Hungarian Rovash Script

The Szekely-Hungarian Rovash script belongs to Hungarian writings on one hand and to the Rovash script (a writing family) on the other hand. The Rovash script indicates a sound according to alphabetic principles: each sound generally corresponds to one letter. In most cases, the line proceeds from right to left, but rarely it is possible to see ancient texts where it can proceed from left to right. In the latter case, the graphemes reflect vertically. There was no distinction made between lowercase and uppercase letters, although sometimes the first letter of proper names was written using capital letter.