Multi-character Field Recognition for Arabic and Chinese Handwriting

Daniel Lopresti\textsuperscript{1}, George Nagy\textsuperscript{2}, Sharad Seth\textsuperscript{3}, and Xiaoli Zhang\textsuperscript{2}

\textsuperscript{1} Lehigh University, Bethlehem, PA 18015, USA
\texttt{lopresti@cse.lehigh.edu}

\textsuperscript{2} Rensselaer Polytechnic Institute, Troy, NY 12180, USA
\texttt{\{nagy, zhangxl\}@ecse.rpi.edu}

\textsuperscript{3} University of Nebraska, Lincoln, NE 68588, USA
\texttt{seth@cse.unl.edu}

Abstract. Two methods, Symbolic Indirect Correlation (SIC) and Style Constrained Classification (SCC), are proposed for recognizing handwritten Arabic and Chinese words and phrases. SIC reassembles variable-length segments of an unknown query that match similar segments of labeled reference words. Recognition is based on the correspondence between the order of the feature vectors and of the lexical transcript in both the query and the references. SIC implicitly incorporates language context in the form of letter n-grams. SCC is based on the notion that the style (distortion or noise) of a character is a good predictor of the distortions arising in other characters, even of a different class, from the same source. It is adaptive in the sense that, with a long-enough field, its accuracy converges to that of a style-specific classifier trained on the writer of the unknown query. Neither SIC nor SCC requires the query words to appear among the references.

1 Introduction

From the perspective of character recognition, Arabic and Chinese are situated at the opposite ends of the spectrum. The former has a small alphabet with word-position dependent allographs, is quasi-cursive, and has diacritics, ascenders, and descenders. The latter has an indefinitely large number of classes (of which only the first $\sim 20,000$ have been coded), essentially word-level symbols (many with a radical-based substructure), and fixed-pitch block characters. Arabic strokes can be approximated by arcs of circles, while most Chinese strokes are straight, with a $\sim 1:7$ range in width (like brush strokes), and a flourish at the end. Unlike Arabic, Chinese does not have deliberate loops.

They also exhibit some commonalities. Both have been incorporated in the scripts used by other languages: Arabic in Urdu and Persian, Han in Japanese and Hangul, among many others. Both have traditional roots and forms dating back several thousand years, preserved in a large body of classical manuscripts, and have undergone considerable and diverse modifications in each host language and region of the world. Nevertheless, both scripts have preserved sufficient
uniformity to link cultures that can no longer understand each other’s speech. Their classical forms are prized and cultivated in calligraphy, which combines visual and language arts. Neither script has upper and lower case.

Industrial-strength Arabic and Chinese OCR products must also be able to recognize Latin characters, “Arabic” (Indian) numerals, and Western punctuation. This introduces additional complexity, more because of the need to handle diverse, intermingled reading orders and output codes than because of the increased number of classes.

Many thousands of papers (the very first of which, coincidentally, is [1]) have been written on Chinese character recognition. By the time of our first survey [2] much of the research was appearing in Chinese and Japanese publications. In our second survey [3], we found little new research in the West. Recent research collaboration with Professor C-L Liu at the Pattern Recognition Laboratory of the Chinese Academy of Science (CASIA), visits with Professor X. Ding at Tsinghua University, and a tour of Hanwang High Technology in Beijing acquainted us with the largest concentrations of character research activity in the world and some of China’s thriving OCR industry.

Research on Arabic character recognition (actually Farsi) began in the late sixties. Scattered projects, mainly by speakers of Arabic in the West, increased until the turn of the millennium, when research began to grow exponentially. Nearly one thousand reports have already been published, mostly in English and French. Nevertheless, work on Arabic OCR lags far behind Chinese OCR because of the lack of monolithic government and market support, and of large, publicly available databases. For a recent survey of the state-of-the-art in offline Arabic handwriting recognition, please see [4].

Extrapolating successful methods from Western (including Russian) OCR is insufficient for either Arabic or Chinese because, ideally, every glyph of an entire document must be considered simultaneously before a label is assigned to any one of them. In practice, this notion translates to field classification, where glyphs that are difficult to recognize in isolation (or that cannot be isolated/segmented) are recognized in conjunction with several others.

Given the wide range of different problems exhibited by the two scripts, tackling both simultaneously is a strategy for research that can bring benefits not only to character recognition on other scripts (such as those derived from Sanskrit), but also to the wider field of pattern recognition. It will foster the development of large vocabulary classifiers that span complex character shapes (Chinese) and complex word shapes (Arabic). Below, we outline how we propose to apply field classifiers, which have already proven successful on easier tasks, to Arabic and Chinese documents.

We elaborate on two orthogonal ideas: Symbolic Indirect Correlation (SIC) and Style Constrained Classification (SCC). The former recognizes unknown sequences of features (possibly spanning several characters) by finding and re-assembling its constituent subsequences in the feature sequence representation of labeled reference text. The unknown word(s) need not be represented in the reference set, only their lexical constituents (i.e., symbol polygrams). Style-based