OPTICAL CHARACTER RECOGNITION OF TOUCHING CHARACTERS

S. Shlien and K. Kubota[1]

Communications Research Centre
3701 Carling Avenue
Ottawa, Ontario, Canada
K2H 8S2

[1]Visiting scientist from Nippon Telegraph and Telephone
Yokosuka Electrical Communication Laboratory,
Yokosuka-shi, 238 Japan

ABSTRACT

A method for isolating and recognizing characters in typeset text containing touching or overlapping characters is described. The method relies on vector quantization techniques to represent the text by an ordered sequence of codes. Characters or character fragments are extracted from this code using a form of string matching and network searches. Compatibility relationships are applied to these fragments to obtain a list of possible input strings. Our simulations on the Symbolics 3640 workstation do not indicate that this method would replace conventional template matching schemes.

KEYWORDS: optical character recognition, template matching, vector quantization, consistent labeling problem.

INTRODUCTION

With the continuous drop in the price of computer memory and VLSI components, it is easy to envision new and more powerful OCR devices becoming available. Such devices would utilize several different recognition strategies depending upon the quality of the input; they would be capable of learning a new font without human supervision; and they would contain large spelling dictionaries for applying contextual postprocessing [1-4]. Software development costs rather than the cost of hardware components would be the limiting factor to building such a system.

Automatic reading machines can process good quality typescript input at error rates approaching those of the average typist, however, only a few machines can handle typeset material containing proportionally-spaced characters [5]. A recurrent problem is the presence of touching characters which invariably results in segmentation and misclassification errors (eg. rn confused with m). Without gaps separating characters, the character extraction and classification process can no longer be performed independently, and this increases the complexity of the recognition problem considerably.

There have been several studies [6-8] on the segmentation problem of touching characters. Some excellent results have been obtained using a combination of dynamic programming and template matching [8]. A new and promising approach uses word shape matching [9]. Our interest in this problem has been motivated primarily by its similarity to the connected speech recognition problem.

In fluent spoken speech, there are no silence gaps between words [10]. The computational requirements for segmenting speech into words presently precludes the operation of any real time speech recognizer on a vocabulary of 5000 words [11].

In both speech recognition and optical character recognition, template matching techniques have been the preferred approach in commercial devices [12 and 13]. Though it is recognized that feature extraction techniques are more robust in handling multifonts, template matching techniques are fast and easy to implement and can handle poor quality input containing noise, voids and touching characters [13]. However, template matching schemes have little tolerance to minor type-face variations, of which there are more than 3000 in common usage [5]. Some of the newer schemes [14] attempt to apply combinations of these two techniques and hopefully reap the benefits of both methods.

Preprocessing techniques in speech recognition rely on data compression techniques (eg. linear predictive coding) in order to extract the essential information from the input data. Some of the more recent recognition methods also apply the rediscovered vector quantization coding schemes [15 and 16] to reduce the data to an input stream of symbols. The Hidden Markov Model [17-21]