Applying Feature Based Word Recognition Approach to Screen Text Recognition

G. Raza, A. Hennig, N. Sherkat, R. J. Whitrow
Department of Computing, The Nottingham Trent University,
Burton Street, Nottingham NG1 4BU, UK
{ghr,amr,ns,rjw}@doc.ntu.ac.uk

Abstract

In earlier work we described a feature based word recognition method for the recognition of poor quality words taken from fax messages. Various independent and robust features are used to identify alternatives of every object of the word without attempting to segment touching objects. Later, a lexical lookup method is used to verify the alternatives In this paper, the developed method is applied to screen text of different fonts and sizes in order to observe its performance on screen image. It has been observed that the developed method is capable of recognising screen text of varying fonts and sizes whilst avoiding segmentation of touching characters.

1 Introduction

Character segmentation is a key step in most conventional Optical Character Recognition (OCR) systems. This segmentation based approach is possible for good quality prints, where characters are clearly separated from their neighbours. However, such an approach is unsuitable for the documents containing characters touching their neighbouring characters. The performance of existing OCR systems is not acceptable for such documents. Documents of this type come from many different sources such as facsimile messages, low quality prints, photocopies etc. Furthermore, existing OCR systems are unable to deal with screen images[1][2].

A feature based word recognition method [3] for the recognition of poor quality word images has been developed. For this paper, this method has been modified and applied to screen text of different fonts and sizes. The results obtained show the ability of the method to cope with screen images.

In this method, different independent features of each object (which could be a single character or several touching ones) of a word are extracted from the input sample word. Different alternatives for each object are found by comparing the features with a database of ideal features for each object. These alternatives are ranked according to the number of features matched. The alternative with the most features matched is considered to be the most likely candidate.

These alternatives are combined to form different words using a dictionary lookup step. These words are also ranked according to the total number of features matched. The word with most features matched is deemed to be the recognised sample word. If there is more than one word, then higher level linguistic information, such as language syntax and semantics, can be used to identify the correct word [4][5][6].

2 Feature Extraction

Feature extraction is an important stage in the recognition of characters particularly in the case of poor quality documents. In order to recognise a character, different features are extracted, which will (hopefully) exhibit the distinctive characteristics of the character[7]. Ideally, the features should enable the recognizer to discriminate correctly betwwen distinct classes of characters.

In the literature different commonly used feature extraction methods have been described , e.g. global features, distribution of points, geometric and topological features, linguistic descriptions, use of context and fuzzy sets. There is no general technique for the design of feature extraction which utilises the designer's a priori knowledge of the recognition problem. Geometrical and topological features are commonly used by human beings in the recognition of patterns because such features can easily be detected by the human eye.

In the current research, various independent features of every object of the sample words are extracted and compared against the ideal form of the object (see Figure 1). Features are: the zones in which the objects are found (middle zone only, upper or lower zones and full, i.e. both upper and lower zone), vertical bars, holes (in the upper,