Character Recognition: SSPR'90 Working Group Report

Thomas Bayer[1], Jonathan Hull[2], and George Nagy[3]

[1] Daimler-Benz Research Center, Wilhelm-Runge-Strasse 11, D-7900 Ulm, Germany
[2] Department of Computer Science, 226 Bell Hall, SUNY at Buffalo, Amherst, NY 14260, USA
[3] Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180-3590, USA

This report summarizes the discussions of the Working Group on Character Recognition of the IAPR 1990 Workshop on Syntactic and Structural Pattern Recognition, Murray Hill, NJ, 13-15 June 1990. The participants were: H. Baird, T. Bayer, H. Fujisawa, T. K. Ho, J. Hull, T. Itagaki, D. Lee, S. Liebowitz, O. Matan, G. Nagy, T. Pavlidis, and S. Srihari. George Nagy moderated the discussion and Thomas Bayer prepared this report based on notes by Nagy, Jonathan Hull, and himself.

It was not easy to agree that any problem in this field is definitely solved. Perhaps this is one: on a small number of known fonts of the printed Latin alphabet, at bodytext sizes ($\geq$ 8 point), under moderate distortions, and in a controlled environment, it should be possible to achieve better than 99.9% top-choice accuracy using a variety of well-understood techniques including dictionary context. However, it's important not to forget that even a 99.99% recognition rate is still unacceptably low for many applications.

It was considerably easier to draw up a list of open problems.

1 Digitizing Resolution

Postal OCR systems use a resolution of 200 ppi; however, 300 ppi seems to have become the standard in office applications. Quite obviously, you need more pixels per inch (ppi), if you change to smaller fonts. Some participants felt that to read 8 point type very well, at least 400 ppi is necessary. In addition to distortions that occur at low resolution, characters will also tend to merge more often, triggering a cascade of problems for later stages. According to studies by Prof. Pavlidis [CVPR'86] OCR recognition rates increase up to 500 ppi, depending of course somewhat on the type of recognition algorithm. When using template matching, performance can actually decrease with increasing resolution, since the high resolution exaggerates features that are detrimental to matching.

2 Sample Image Databases

OCR research is strongly influenced by the availability of large databases of scanned text. It was pointed out that the IEEE has recently begun a pilot project

appeared in "Structured Document Image Analysis," H.S. Baird, H. Bunke, and K. Yamamoto (eds.), Springer Verlag, Berlin, 1992, pp. 565-567.

to disseminate their publications in electronic form. This is done by scanning pages at 300 ppi and storing the binarized image in compressed form on CD-Rom. This database may provide an important experimental environment as well as an application domain for OCR research.

3 Distortion Modeling

A common technique for adapting recognition algorithms to distorted patterns is to train the system on large sample sets. However, this technique is time consuming and expensive since the patterns must be collected and each pattern must be labeled. An alternative technique is to analytically model noise so that, starting from a prototype of each class, an arbitrarily large set of samples can be inferred automatically. The advantage is obvious: no time consuming collection of a sample set and a very easy labeling process. However, both techniques have to model a truly representative set of actual samples, else the methodology is suspect — and it is often difficult to make convincing arguments that any given set is or is not representative.

4 Non-zero Skew Angles

The existence of even small amounts of skew (rotation) can be troublesome to text recognition algorithms. In commercial systems this is currently a severe problem. In handwriting recognition skew normalization is as necessary as normalization of size and aspect ratio. Some approaches are available to de-skew a complete document. However, they only work successfully in specialized applications, such as accurately printed documents where there a single dominant skew orientation.

5 Style Consistency

The detection of the font style, point size, etc. of a text is an obvious way to improve the capabilities of text recognition algorithms. This would allow for hundreds of fonts to be used for training but retain the recognition accuracy and potential speed of a system that uses a small number of fonts. This appears to be a promising but hitherto almost neglected topic.

6 Locally Adaptive Algorithms

Text recognition algorithms that can be readily (ideally fully automatically) adapted to specific constraints are interesting. For example, say you have a recognition system trained on three fonts. If a fourth font is added to the system, recognition performance should not decrease. So far, there is little to nothing published on this issue.

7 Contextual Information

This subject has attracted attention for quite some time, and many useful techniques have been developed. However, specific techniques may require information not provided by current methods. For example, syntactic analysis of languages has been shown useful in some domains. However, only a few syntactic models have been exploited for OCR. The utilization of semantics of languages has been even more limited.

8 Feature Selection

The general problem of feature subset selection is well known in pattern recognition. An area of current interest not covered by most classic techniques is the dynamic selection of the most useful features based on the characteristics of an image.

9 Combination of Recognition Techniques

It has been observed that a series of classifiers each based on a specific limited feature set can be applied in parallel to yield performance that is potentially better than using one classifier based on all the features. An open problem is how to effect the combination to achieve performance consistently superior to all the individual classifiers, for any sufficiently rich mixture of classifiers.

10 Should 100% Top-Choice Accuracy Be The Goal?

On isolated characters, it may not be possible efficiently to achieve close to 100% correct recognition rate if noise introduced by scanning and preprocessing is too great. Such noise can degrade even human performance to 80%. However, under the same conditions, people can achieve 100% correct when word context is provided. Therefore, it might be sufficient for isolated-character recognition algorithms to exhibit a peak correct recognition rate of only, say, 80% on such data, so long as it achieves much higher accuracy *in top* k, for small $k > 1$ (or some similar measure), so that the remaining errors can be corrected by contextual postprocessing.