A robust preclassification scheme for hierarchical character recognition

A Robust PreClassification Scheme For Hierarchical Character Recognition

N.A. Khan, J.A. Hegt, I. Cebollero
Eindhoven University of Technology
Dept. of Electrical Engineering
P.O. Box 513, 5600MB Eindhoven, The Netherlands
E-mail: n.a.khan@ele.tue.nl

Abstract—**This paper presents a structural coarse classification scheme to produce a small set of candidate character classes that are interesting to be considered for matching the given sample at the main classification stage. It is build as part of a hierarchical recognition scheme to increase the overall systemspeed and to improve the recognition accuracy. The approach is based on preparing highlevel coarse shape models of character classes permitting similar characterclasses to merge into super groups. The approach uses minimal number of models and is robust to distortion and font or writing style variations.**

Keywords—**Coarse shape description, Character recognition, Hierarchical classification, Handwritten numerals**

I. Introduction

A pre-classifier that can quickly screen out character-classes that are very unlikely to match a given input sample of character (hand-written or multi-font machine printed) can prove to be useful in a number of ways [1], [2]. For example:

1. To accelerate classification speed by avoiding the detailed examination at the main classifier stage for unlikely character-classes
2. To obtain high and reliable recognition specially in cases of large number of character-classes like Chinese characters. For instance, in such cases when neural networks are employed, problems like very slow learning speed and difficulty in reaching convergence due to ”local minima problem” are encountered [1].
3. To aid in word-segmentation for producing valid character hypothesis for recognition.

On the other hand, it is important that when dealing with distorted samples, the pre-classifier does not breakdown but gracefully decreases in its prototype screening efficiency. Moreover, the focus should also be to keep the number of reference models (prototypes) minimal for high speed.

A scheme is proposed keeping in view above goals. Though the focus is on upper case English alphabet and numerals, the approach is certainly not restricted to it.

II. The Approach Outline

The basis of the approach is to prepare coarse structural (or shape) definitions of character classes permitting similar character classes to share the same or similar models. The aim is that the screening efficiency or the output set produced by the pre-classifier should be related to:

a Resemblance of the input sample to the defined super-class shape models(s);
b Similarity in shape of the input character class to other classes; vague shapes of the input sample generally triggering a larger output set;
c The distortion level expected (or allowed) in the input samples.

In our approach, these coarse definitions of the character classes are effectively derived by defining a set of ”super-feature” families coarsely modelling parts of a character body and representing it as a class structural model. We will refer to them as (super-) class prototypes. The classification of an input sample is performed by first extracting a set of features that can match the employed super-features and checking for their coarse spatial relation

III. Super-Features and Class Prototype Modelling

The scheme employs three family types of super- features. Their definitions in terms of what parts of a character they can coarsely model and represent are as follows:

Loop Any bounded or closed convex polygonal part of the character body. Its centroid is used as the reference point for the rough relative spatial location determination in a character.

Bay family Any unbounded convex polygonal part of the character body such that the angle between the first and last line-segment or stroke is below 90 °. The family includes eight bays in eight principle directions: North, North-West, West, South-West, South, South-East, East, North-East. Here too, its centroid is used as the reference point for determining the rough relative spatial location in a character. Thus bays found in a character-

ISBN: 90-73461-15-4

body are quantized in one of these eight principle directions keeping in view the variation range in the bay's direction.

Line-segment family Any straight line-segment of the character body which cannot be included in a bay or loop. This family comprises of line-segments with four principle orientations: vertical, horizontal and 45 ° inclined to the horizontal. Once again, its centroid serves as the reference point for determining its rough spatial location in the character

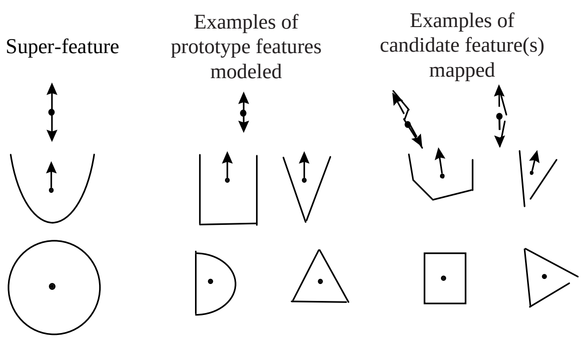
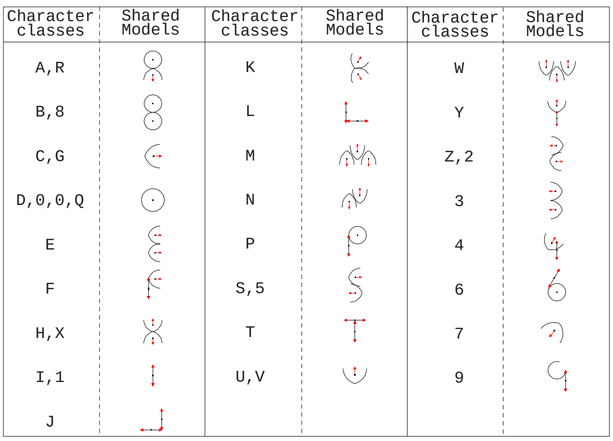


Fig. 1. Some super-features with examples of character parts that can match them in cases of super-class prototypes and the input samples

This proposed mapping of super-features to prototype parts is illustrated in fig. 1 with few examples. Finally, the proposed class prototypes resulting from this kind of modelling are shown in fig. 2 as graphs of super-features. For computation ease each prototype is stored as a vector of super-features with separate fields for all the pertinent information. Additionally, a relation matrix coarsely specifying the relative angular position (in 8 principal directions) of every super-feature w.r.t. the others is also stored. Note that generally two super- features are involved in constructing a prototype allowing quick processing.



Character classes	Shared Models	Character classes	Shared Models	Character classes	Shared Models
A,R		K		W	
B,8		L		Y	
C,G		M		Z,2	
D,0,0,Q		N		3	
E		P		4	
F		S,5		6	
H,X		T		7	
I,1		U,V		9	
J					

Fig. 2. Proposed set of super-class prototypes represented as feature graphs

IV. Superfeature Detection in the Input Sample

The input image of the character is pre-processed (for noise removal and enhancement) and then thinned. After size normalisation and removal of spurious segments, it is coarsely polyline fitted with straight line segments. For better matching, vertices are also added at the junctions, crossings and where the end-point of a line segment is in close proximity to another line segment. Small line-segments corresponding to serifs are rejected.

All possible loops, bays and line-segments are extracted from the character polygram and the feature vector is derived which we will refer as the candidate super- feature vector. The candidate super-feature vector differs from the prototype one that the actual directions of the bays and the orientations are not quantized but kept as an angle value.

When identifying candidate super-features, somewhat loser definitions are used to accommodate the non- idealities. This is illustrated also with examples in fig. 1. For instance, small gaps are permitted. In case of linesegments an adjacent set of polylines can be grouped to be seen as a straight line-segment super-features, provided the deviation from straightness do not exceed a certain limit. In case of a bay, the permitted angle limit between the ends is larger (e.g 100 °). Note that multiple interpretations are permitted of the same or overlapping parts of the polygram. That is appropriately screened out later in the context of the prototypes.

V. The Screening Process

A prototype is included in the output list if the following three conditions are satisfied:

(i) If all corresponding prototype super-features can be found in the candidate super-feature vector . For this purpose, flexible matching between the candidate and prototype super-features is permitted. A candidate super-feature can be semantically equated to a same family type prototype super-feature allowing a maximum of 67.5 °[1] orientation difference.

(ii) If a similar basic relative spatial inter-relation or arrangement holds among the identified candidate super-features as depicted by the prototype superfeatures. For that purpose, following is checked to hold for all the corresponding pairs of super-features in the prototype and the candidate:

- in case an ordinate of one of the prototype superfeature is greater (or smaller) than the ordinate of the other prototype super-feature the same should hold in the candidate pair.

[1] Selected value based on trial and error

- in case the two prototype super-features have equal ordinates, the ordinates of the candidate super-features can have a relative displacement of no more than half the length (or width) in the perpendicular direction of either of the super-feature involved.

(iii) Finally, the total height and width of the candidate is checked which should not be less than a certain limit (e.g. two-third) of the reference size of the prototype. This helps to eliminate prototypes which inherently possess a subset of the super-features of the other prototypes like a loop of a ”0” is found in a ”8”.

If the candidate super-feature vector is larger than the prototype super-feature, multiple options for matching are tried so that all options are exploited to see if the desired matching can be established. Fig. 3 provides some illustration of the application of the scheme in few cases.

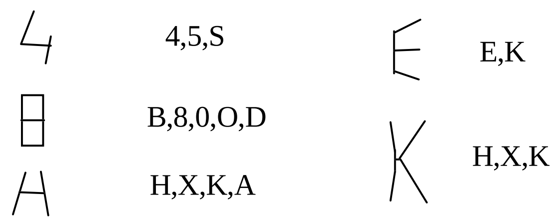


Fig. 3. Examples illustrating the results of pre-classification scheme in various cases

VI. Experimental Results

Though the system is still under improvement to be followed by an extensive performance evaluation, to date results are already promising. When tried on a database comprising of few hundreds of pre-segmented totally unconstrained but without gaps hand-written alphanumeric characters from CEDAR CDROM-I database that are extracted from real-life mail-pieces, a reduction rate of 85.7% in the character-classes to be compared at the main classification stage has been achieved. Most important of all is that the correct prototype is always in the output set. The database includes examples as shown in fig. 4. The system can cope with segmentation errors, extraneous strokes, distortions introduced during stroke-extraction stages etc. However, the ability to handle gaps still has to be built in .

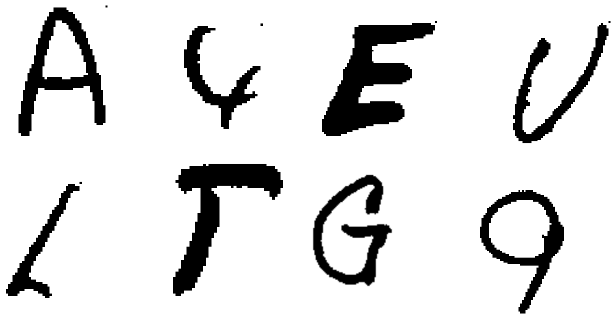


Fig. 4. Some examples of real-life images of hand-written alphanumerics correctly pre-classified by the system

VII. Concluding Remarks

The proposed pre-classifier scheme utilises high-level pre-prepared coarse descriptions of character-classes (or super-class models) using a set of ”super-features”. The scheme lends itself to a hierarchical classification scheme more closer to human aesthetic sense that can be employed for a high and reliable recognition performance [3]. The approach does not breakdown with distortions but gracefully decreases in the prototype screening efficiency as the distortion level increases. However, for an optimal screening efficiency and speed, the allowed flexibility in mapping parts of the input character's skeletal polygram should be tuned according to the expected distortion levels.

References

[1] Y. Waizumi et.al., ”High Speed Rough Classification for Handwritten Characters Using Hierarchical Learning Vector Quantization”, *Proc. Fourth International Conference on Document Analysis and Recognition*, Vol. 1, pp 23-27, Aug. 1997

[2] A. Marcelli et. al, ”Structural Indexing for Character Recognition”, *Computer Vision and Image Understanding*, Vol. 66, No. 3, pp 330-346, June 1997.

[3] N.A.Khan, H.A.Hegt ”An Optimal Shape Modelling Scheme for Omnifont and Handwritten Character Recognition”, *ASCI'98*, pp 92-97, June 9-11 1998.

This page was intentionally left blank