NUMBER OF FEATURES FOR HANDWRITTEN CHARACTER RECOGNITION

R.Krishnan Tampi[1] C.S.Sridhar[2]

[1] F.A.C.T.Ltd, Udyogamandal, Cochin, India
[2] Cochin University of Science and Technology, Cochin - 682022, India

RÉSUME

SUMMARY

Definition of features for general class of images is a difficult problem. For a particular class -- handwritten lower case English script -- it is shown that features can be defined on an arithmetic basis. This is achieved by dividing individual characters into fixed number of equal areas and using a finite set of primitives to define a best approximation to the contour existing in that area.

Introduction

Handwriting recognition requires the correct description of images and appropriate definition of features. Suen recommends a 30x40 pixel size for the recognition of binary images and alpha-numeric characters [1]. However, feature ectraction remains a problem with no simple solution in sight. A recogniser which uses structural and global properties would be the most appropriate approach to the extraction of topological features. In any case computation required will be enormous. A question arises whether for a small set of characters which have to conform to certain universal shapes such a complex approach is essential.

It is common knowledge that highly distorted, contour approximated letters in handwriting are easily recognised by human beings. Acknowledging the fact that humans have a large amount of learning the result remains unaltered that, in the recognition domain these characters are seperated by distances which are large and are invariant to the type of algorithm used by recognisers. In this paper it is shown that features can be defined on arithmetic basis for use in recognition algorithms. 36 features and two auxiliary are defined to assist recognition.

Feature definition on arithmetic basis

Consider the image of a single character. Let 144 pixels of its binary image arranged in 12 rows and 12 columns describe it completly. Let this be divided into 16 subregions called cells. Let each cell contain 9 pixels in 3 rows and 3 columns. This ceil is used to define features.

The contour which passes through each pixel is binarised, its presence denoted by a **1** and absence by a **0**. Thus the image is entirely represented by these 9 pixel square cells. This can be transformed to a 9 bit binary word with the central pixel bit as the MSB as shown in Fig.1. After analysing a large number of samples a few unique contours were selected as features for this class of images.

Since each cell has 9 elements there could exist 512 features provided they are unique. In natural writing all these do not exist and broken bits, tails and such artifacts cannot be treated as features. Looking at the binary words whcih represent the contours, we can find that not all of these are unique, being only rotations of other words. It can then be implied that such words which are mere rotations of others cannot be qualified as features.

Fig.2 shows two examples of describing the image of a character in the 16 cell paragraph format. 144 pixels have been divided into cells of 9 elements and also shown is the method of deriving the 9 bit binary word from the cells. Consider now a single cell. If it is rotated without altering its position in the paragraph, the central bit remains unshifted in position while the other 8 undergo shift. But the bit pattern remains unchanged. This is clearly an arithmetic operation on the 8 LSB's of the word. Each possible word which represents a contour can therefore have 7 rotations with reference to itself. Out of these eight only one can be unique. These unique words are to be considered as features. Conversely a feature can be defined as one which is not derived from another. Thus if a binary word representing a contour cannot be derived by lower order rotation from another such a word can be treated as a feature.

The number of features for this type of a description of an image can now be calculated by calculating the number of unique allowable 8 bit words. Such words are called Basic Words. A basic word must be odd decimal, and they can be listed out with mod 64 logic. It is also possible to arrive at a basic word from any given word. Consider the word say, **0 10 110 100.** (For simplicity octal notation can be used). The word and its rotations are, 264, 151, 322, 245, 113, 226, 055 and 132, the MSB not considered since it does not undergo rotation. Amongst all these, 055 is alone unique and is the basic word. 055 is the smallest number and all the rest can be derived from it. A simple calculation shows that there are only 36 such basic numbers (8 bit words suitable for character cells) with the 9th bit 0 and an equal number with the 9th bit 1. A list of the basic words is given in Table-I. All the remaining words from the 512 possible can be generated from these words.

For facilitating computer programme development, these basic words are given a 3 digit name. The first digit indicates the number of 1's in the word, the second indicates the family of the contour, but normally chosen so that the smallest basic

corresponds to the smallest 2nd digit of the name and the third digit representing the rotation with respect to the basic word. Some of the basic words have only four unique rotations while some others have none. Thus a total 512 are accounted for. For the type of image, cursive script handled here, three basic words are illegal as they represent broken contours. Considering all these factors 68 features can be defined.

Sufficiency of the features defined -

The 26 characters, the targets for classification in this work, show 0.6 as the cell occupancy and 0.25 as the pixel occupancy rate. The features defined above are sufficient for the classification of these images since the rotational forms can take care of any contour that may appear in handwriting. However, this is a very large collection of words which will be difficult to implement on small machines. Hence data reduction is adopted. Data reduction, called filtering here, takes the form of replacing all words with more than 3 1's by words with 3 1's. Eight words are retained, which when used for reconstruction of the character show remarkably good images. A further reduction was attempted, wherein only 4 words are retained. These are symmetric horizontal, vertical, right and left slants. Fig.3 illustrates the results of filtering for some character images. The retention of 'letter' shape after the first filtering (8 words) is good. Also remarkable is the fact that when only four words are used the reconstructed images are highly distorted no doubt, but, the image of one character does in no way resemble that of another. This gives a good scope for classification.

However, with only four features available syntactic methods for classification lead to a poor recognition score. This can be offset by a stricter replacement method during filtering. A preservation measure is defined to ensure stricter replacement. The trajectory and the continuity of the contour have to be used as measures for good shape retention. Therefore two secondary features are defined. These are called Tense and Tendency. They do not have the status of features. The extraction of these auxiliary features is fairly straight forward. The last digit of the word stands for the tendecncy, while the last two digits indicate the tense. Fig.3 illustrates these aspects. It is to enable the extraction of features like these that the word names for the basic words were appropriately chosen. With the paragraph of 16 cells and 4 features supported by auxiliary features recognition is possible even using small macines.

Conclusion -

It is shown that for lower case English (Roman), handwritten characters a small number of features can be defined on an arithmetic basis. Using auxiliary features recognition can be implemented based on four features. Though only handwritten characters were considered, this approach may hold good for similar character sets and images which have a universal quality.

Reference -

[1] C.Y.Suen, "Character recognition by computer", Handbook of Pattern Recognition, Eds. K.S.Fu and T.Y.Young, Academic Press, N.Y., 1986.

Number of Features for Handwritten Character Recognition

Table I
Basic words

Octal Word	Word Name	Word Name MSB=0	Word Name MSB=1	Comment
000	000	000	000	Null
001	001	001	001	Illegal
003	200	200	307	Legal
005	210	210	380	
007	300	300	490	
011	220	220	390	
013	310	310	4A0	

and similarly the following in octal
015, 017, 021, 023, 025, 027, 031, 033
035, 037, 045, 047, 053, 055, 057, 063
065, 067, 073, 075, 077, 125, 127, 133
137, 157, 167, 177 and 377

```
000000000000
000001100000
000011100000
000011100000
000011000000   000
000111000000   001
000110000000   001   --    0 00 011 000
001100011110
001100110000
001100010000   110
010111100000   100   --    0 11 000 011
000000000000   100
```

Binary image ". Example of cells and their binary words.

Fig.1 A character image, the cells at position 31,32 and the binary words.

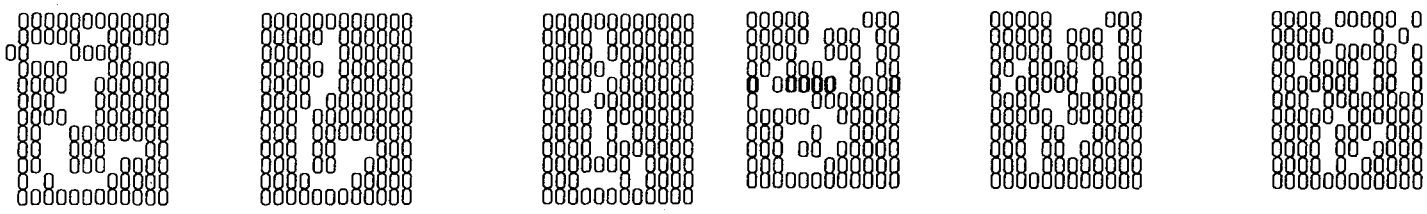
```
596 400 405 000      000 402 300 306
306 000 102 206      426 000 582 323
306 000 203 306      201 600 405 000
4E4 304 382 3A0       000 604 316 000
```

paragraph descrption of 'a' and 'd'.

```
  100              110        111
1 100   after one 100 2nd rot 000
  100    rotation 000 ation 000
```

Cell 306 and its rotation. Names are 307 and 300.

Fig.2 Paragraph of two characters and example of rotation.



'b' Filter I Filter II 'g' Filter I Filter II

Fig.3 Effect of filtering 'b' and 'g'. 1's are suppressed.