CASE STUDY ON TAMIL CHARACTER ANALYSIS USING IMAGE PROCESSING

[1.] Dr.Belsam Jeba Ananth.M, [2.]Vennila.M

[1.] Professor (EEE) / DMI College of Engineering, Chennai

[2.] Associate Professor (EEE) / DMI College of Engineering, Chennai

E-mail:dmice.eee@gmail.com

Abstract: -

Handwriting Analysis has been a popular subject of research due to its vast applications in fields like signature analysis, address reading, and converting written text into hyper text etc.There have been several complexities that arise during written letter analysis especially when there are no constraints on the text to be analyzed. This paper presents new techniques for slant and slope removal in cursive handwritten words. Both methods require neither heuristics nor parameter tuning. This avoids the heavy experimental effort required to find the optimal configuration of a parameter set. The proposed technique is shown to improve the recognition rate by 10.8 % relative to traditional normalization methods. Moreover a long exploration of the parameter space is avoided. Any piece of handwriting is written at a certain slant. Slant is one of the characteristics that make handwriting harder to process automatically than printed text. For this reason slant correction is a standard step in systems for processing written text. Slant and Slope can be introduced by both handwriting style and acquisition process. Ideally the removal results in a word image independent with respect to such factors. This process is called normalization. The normalization aims to transform the handwritten data into segments. Our slant correction method uses the vertical projection histogram .The idea is that the histogram of a word that is written straight up will have larger and more distinct peakes.Therefore we could look at the histogram of the word at different shear angles and take the one with the highest peak.

Index Terms: Tamil Characters, Support vector machine, Preprossing, Feature extraction and Slant Removal.

Introduction: Handwritten character recognition is a challenging problem in pattern recognition area. The difficulty is mainly caused by the large variations of individual writing style. Robust feature extraction is very important to improve the performance of handwritten character recognition system. In this paper we proposed a combined feature based on the gradient feature and coefficients of wavelet transform. Generally speaking, the gradient feature represents local characteristics of a character image properly, but it is sensitive to the deformation of handwritten character. Meanwhile, wavelet transform represents the character image in multiresolution analysis and keeps adequate global characteristics of a character image in different scales. In order to improve the discrimination of a feature, it is better to compose local and global characteristics into a combined feature. We conducted experiments on Tamil alphabets, to evaluate the performance of our feature.

Slope and Slant Removal: This paper presents new techniques for slant and slope removal in cursive handwritten words. Both methods require neither heuristics nor parameter tuning. This avoids the heavy experimental effort required to find the optimal configuration of a parameter set. A comparison between the new de-slanting technique and the method proposed by Bozinovic and Sridhar was made by measuring the performance of both methods within a word recognition system tested on different databases. The proposed technique is shown to improve the recognition rate by 10.8% relative to traditional normalization methods.

Cursive Script Recognition: Off-line cursive script recognition (CSR) is an important technology for applications involving automated processing of hand written data, such as bank check processing and postal address reading. Almost every CSR system presented in the literature involves a normalization step, which consists in removing slant and slope. The slant is the angle between the vertical direction and the direction of the strokes that, in an ideal model of handwriting, are supposed to be vertical. The slope is the angle between the horizontal direction and the direction of the line on which the word. Slant and slope can be introduced by both handwriting style and acquisition process. Ideally the removal results in a word image independent with respect to such factors. For this reason the process is called normalization. The recognition techniques usually applied Dynamic Programming (DP), and Hidden Markov Models (HMMs) need a fragmentation of the word. In DP based systems where the isolated submits are expected to be characters, the normalization reduces the shape variability, which simplifies their classification by pattern recognition techniques. In HMM-based systems, where the signal is assumed to be piece-wise stationary, the normalization aims to transform the handwritten data into such segments.

In the literature the description of the normalization technique is often neglected and the use of heuristic rules, frequently based on parameters that need to be set manually is apparent. Parameter tuning can be a time consuming process and only be optional for a specific database.

The work presents new techniques to remove slope and slant that avoid such problems and that do not involve any heuristic the new de-sloping technique, The de-slanting technique is compared with the widely applied “Bozinovic and Sridhar Method” in terms of recognition rate describes a common slope removal algorithm.

Traditional Slope Removal Algorithm: The traditional method for slope removal starts by finding a first rough estimate of the core region, the region enclosing the character bodies. This estimate is biased by the fact that the word is not horizontal, so that the upper and the lower limits of the estimated core region do not fit as they should, the extrema of the character bodies. To solve this problem, the stroke minima closer to the lower limit of the estimated core region are close to fit the line connecting the button points of the character bodies. This is the line on which the word is aligned. The image is de-sloped when this line is horizontal, which is a condition achieved by a rotational transform. After the de-slanting step, the core region can be re-estimated under the assumption that the word image is now horizontal. It becomes evident that the first estimate of the core region is based on the hypothesis that the density of the lines belonging to it is higher than the density of the other lines. This is evident in the horizontal density histogram showing higher values in correspondence with the core region, and this property is used to perform the detection.

Core region lines are usually obtained as the ones surrounding the highest density peaks, but this technique is strongly affected by the presence of long horizontal strokes that can be confused with the actual core region and can lead to severe errors in normalizing the word. Many heuristic rules are necessary to handle the problem and the resulting process is not robust.

Our Slope Removal Algorithm: In order to avoid the problems described in the previous section we analyse the distribution of the density histogram itself. The density distribution P (h) is expected to be bimodal. One mode corresponds to the high density lines of the core region; the other corresponds to the other lines. when the density is low. A threshold can be used to distinguish between core region lines and remaining lines. Since the horizontal strokes have a negligible influence of P (h), the threshold is robust with respect to their presence. The density histogram H(i) presents the density for each row i. The density distribution describes the probability P(h) of having h foreground pixels in a line. Long horizontal strokes give rise to peaks in the density histogram but their influence in the density probability distribution is small because they only occur a few times.

We use the Otsu thresholding algorithm that is based on the minimization of the weighted sum of the variances of the following two sets: the group of the density elements less or equal to the threshold and group of the density element greater than the threshold. The weights are the probabilities of the respective groups, calculated as the percentage of elements belonging to each group. When several regions have a density higher than the threshold, the region with the highest number of foreground pixels is selected as the core region.

The New Deslanting Technique: The new deslanting technique is based on the hypothesis that the word is deslanted when the number of columns containing a continuous stroke is maximum. For each angle is a reasonable interval, a shear transform is applied to the image and the following histogram is calculated. Where is the vertical density in column m, and the distance between the highest and lowest pixel in the same column. If the column m contains a continuous stroke (m)=1 otherwise H(m)=[0,1].For each shear transformed image, the following function. S (α) =∑ hx (i) ^{2} is Calculated. The angle α giving the highest value of S is taken as the slant estimate by using in S (α) the square value of the density rather than the density itself, the contribution of the longest vertical strokes is enhanced with respect to that of short strokes, which are often due to strongly slanted vertical letters. Histogram and function are easier to calculate than the above mentioned angle histogram. The detection of near vertical strokes their borders and their approximation with a set of straight lines strokes are in fact not required. A simple count of the foreground pixels of each column allows as calculate and the need for multiple Shear transforms makes the method computationally heavy but the computation can be reduced using some recursive procedure to obtain the shear transformed image for angle from the shear transformed image for angle. On the other hand the absence of parameters avoids any experimental effort in parameter optimization.

Experiments and Results:

The effectiveness of the proposed methods was evaluated as a function of the performance of a CSR system and compared to traditional normalization methods. The CSR system converts the handwritten data into a sequence of vectors with a sliding window shifting column from left to right. At each window position, a frame is isolated and a feature vector is extracted. The Recognition is performed using continues density Hidden Markov model to calculate the likelihood of the observation sequence with each word in a lexicon. The most likely word is selected as the interpretation of the handwritten data. Training is based on Maximum likelihood estimation; While Recognition is based on the estimation of maximum a posteriori probability.

References:

1. An adaptive Technique for handwritten Tamil Character Recognition-IEEE 2007, Sarveswaran.K and Ratnaweera.D.A.A

2. Enhancing the performance of handwritten Tamil character recognition by slant removal and introducing special features.-Journal of soft computing-2008.N.Shanthi and K.Duraiswamy.

3. Recognition of hand printed Tamil Characters, Patteren Recognition -1980, Chinaswamy and S.G.Krishnamoorthy.

4. Character Recognition a Review-Pattern Recognition 1990, Govindan and A.P.Shivaprasad.

5. A complete OCR system development of Tamil magazine document-2003, Aparna.K.H.