504/E1

Automated response recognition system for questionnaires

M A I D Fernando[1]* and N V Chandrasekara[2]

[1]Department of Statistics, University of Colombo, Colombo 03
[2]Department of Statistics and Computer Science, University of Kelaniya, Kelaniya

An automated system capable of recognizing responses for questionnaires and entering them into the database will be very useful in many subjects. Entering data manually is time consuming. Thus, through this research, a new clustering method to cluster printed and handwritten words, and a character recognition method to identify each character of handwritten words were discovered.

An automated system to recognize response should be capable of separating printed words from the handwritten answers in a questionnaire, and recognizing each character in the word. Horizontal segmentation and vertical segmentation were used to segment the lines of the scanned questionnaire and segment the words in each line respectively. Two types of data, characters including 26 English upper case alphabet characters, 10 numeric characters and 3 main symbols, dot (.), at (@) and dash (-) and words including printed words and handwritten words, are collected using a questionnaire. The target population was students of the University of Colombo, Faculty of Science with a population size of 1500. Stratified sampling is the method which was used to collect data. The sample size was chosen as 300 where the marginal error of the sampling is 0.05. Thus, 16 strata were created by considering the facts gender, stream of study and year of academy.

Six features were identified as height, pixel density, pixel distribution, vertical project variance, major vertical edge and major horizontal project profile, to cluster the printed and handwritten words. The results indicated that agglomerative hierarchical clustering provides the highest recall accuracy of 98%. Complete distance linkage and Euclidean distances maximize the Cophenetic correlation coefficient as 0.8874. Once the handwritten words are recognized, vertical segmentation was used to separate characters of the word. 16 partial densities were calculated for each character as features. Assuming that the large number of data behaves according to a Gaussian distribution, a Probabilistic neural network was created with an input layer which contains 16 partial densities as variables and an output layer which resulted in 39 classes including 26 English upper case characters, 10 numerical characters and 3 symbols. The system shows the recall accuracy as 71.4% when the spread was considered as 14.

The major drawback of the system was the difficulty of separating number 0 and character O, number 1 and character I, number 2 and character Z, number 5 and character S. This reduced the accuracy of character recognition. Still, the system provides a better solution to automate the data entering of a questionnaire by providing great efficiency.

inoshifernando@gmail.com Tel: 0382238348