Automated Response Recognition System for Questionnaires

Inoshika Dilrukshi
University of Colombo School of Computing
inoshi.fernando@gmail.com

N. V. Chandrasekara
Faculty of Science, University of Kelaniya
nvc.stat@gmail.com

Abstract— 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, the purpose of the research is to automate the manual data entry process. Through this research, a new clustering method to cluster printed and handwritten words, and character recognition method to identify each character of handwritten words was discovered. An automated system which recognizes response should be capable of separating printed words from the handwritten answers in a questionnaire, and recognizing each character in the handwritten word. Horizontal project profile was used to segment the lines of the scanned questionnaire and vertical project profile to segment the words in each line. Two types of data, characters and words were collected using a questionnaire. 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, The target population was students of University of Colombo, Faculty of Science with a population size of 2000. Stratified sampling is the method which used to collect data. Sample size was chosen as 300 where the marginal error of the sampling is 0.05. Thus, 16 strata were created considering facts gender, stream of study and year of academy. Six features are 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. Results discovered that agglomerative hierarchical clustering provides highest recall accuracy of 98%. Complete distance linkage and Euclidean distances maximize the Cophenetic correlation coefficient as 0.8874. Once recognize the handwritten words, vertical project profile 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 Gaussian distribution, Probabilistic neural network was created with an input layer which contains 16 partial densities as variables and output layer which results 39 classes including 26 English upper case characters, 10 numerical characters and 3 symbols. System shows the recall accuracy as 71.4% when spread was considered as 14. 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 was a reason to reduce the accuracy of recognizing characters. Still, the system provides a better solution to automate the data entering of a questionnaire by providing high efficiency.

Keywords— Probabilistic Neural Network, Clustering, Text recognition

I. Introduction

In Statistics, the information will be processed using data. There are different types of data collection methods. Conducting a survey is the most popular method among them. Usually, a survey is conducted using questionnaires. In Sri Lanka, paper-based questionnaire is the most popular method of collecting data due to lack of technology resources. Yet, maintaining paper based file system requires more physical storage space with the risk of destruction of information. Saving data electronically in computers reduces the risk of information loss and increase the effectively and efficiency of analyse.

Difficulties of manually data entering to the database are the major problem of using paper based questionnaires as it is time consuming. Tegang et al. [1] states that manual data entry will produce an error of 1%. Thus, the researcher introduced personal digital assistant (PDA) software concept to reduce this error. Still with lack of technology and knowledge, it is unable to use such techniques in Sri Lanka.

The suggested method automates the manual data entry. In this method, data was collected using paper-based questionnaires and each questionnaire was scanned. These scanned images are the inputs for the system. Horizontal project profile was used to separate the lines of the scanned image and vertical project profile was used to separate the words. This results a set of printed words and hand written words. These printed words and the handwritten words are then separated using appropriate techniques. Clustering techniques are used to separate these printed and handwritten words. There are 6 features which were used to conduct the clustering process. Those are height, pixel density, pixel distribution, vertical projection variance, major vertical edge and major horizontal projection difference.

Once the handwritten words are identified, it had been segmented into characters using vertical project profile. These characters were then forward for several pre processing steps which cause to reduce noise data. These processed images can be identifying using the neural network. A probabilistic neural network was used in order to recognize characters. The partial densities were used as features in order to train the PNN network.

Following chapters will describe the process of analyzing. Section 2 describes the methodology of the process. This includes the process of data collection, pre processing, feature selection for clustering and feature

selection for classification. Section 3 describes the analyzing and discussion of the system. Section 4 describes the conclusion of the findings.

II. METHODOLOGY

The system was a process of data collecting, image pre processing, selecting features for both clustering and classification.

A. Data Collection

The suggested method contains basically 2 tasks: obtain a clustering method to separate printed and handwritten words is the 1st task and training a neural network to identify handwritten characters is the 2nd task. In order to fulfil these two tasks, two types of data are required to carry out the research. These two types of data, characters and words, are collected using questionnaires. Both printed and handwritten words are required to develop the clustering method. Characters including 26 upper case alphabet characters, 10 numeric characters and 3 main symbols, dot (.), at (@) and dash (-) are required to develop the neural network in order to identify characters. The target population was University of Colombo, Faculty of Science with size around 2000. From the population, a suitable sample was required to be selected. Thus, the size of the sample was chosen as 300 according to the table, determining minimum returned sample size for given population, created by Bartlett et al. [2]. Samples were created using Stratified sampling. Many characteristics such as gender, age, usage of hand writing will cause to the difference of hand writing from person to person. Thus, 16 strata were defined considering gender, year of academy as 1st year, second year, 3rd year and 4th year and stream of study as Bio and Physical in order to gather different types of characters. The stream of study was used as Bio students are tending to use alphabet characters and physical students tend to use numerical characters. Table 1 shows the size of the each stratum.

B. Image pre processing

In analyzing a questionnaire, the questions and answers should be segment into words. Some researchers perform character segmentation before the actual character recognition while others avoid the segmentation stage. Kavallieratou et. al. [3] and Herath et. al. [4] both state that horizontal projection profile is the most commonly used method for line segmenting and vertical projection profile for word segmenting. These words were clustered and hand written words are chosen in order to recognize characters.

The next task is to pre process the characters which will be use to train the data. The collected images of characters were cropped and binaries as given in Fig. 1. The threshold to binaries the images was calculated using Otsu method thus, da Silva et. al. [5] suggests that Otsu bi-level approach determines the optimal threshold. White space which was surrounded the letter was removed using suitable MATLAB codes. The

TABLE I. SIZE OF EACH STATUM IN TARGET POPULATION

	1st year		2nd year		3rd year		4th year	
	Bio	*Phy*	*Bio*	*Phy*	*Bio*	*Phy*	*Bio*	*Phy*
Male	17	188	64	222	49	163	25	80
Female	125	112	125	124	100	94	77	78

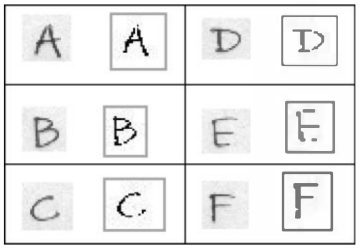

Fig. 1. Binarised images

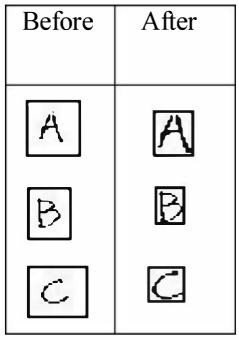

Fig. 2. Images after removing white spaces

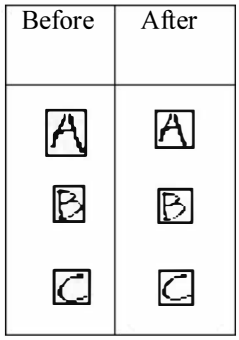

Fig. 3. Scaling and resizing

examples of resulting images were shown in Fig. 2. These resulting character will be scaled and resize into a 16×16 square matrix as given in Fig. 3, in order to use for character recognition network

C. Feature selection for clustering

In order to separate handwritten words and printed words, da Silva et. al. [5] states the most important features as pixel density, vertical projection variance, major horizontal project difference, pixel distribution, vertical edges and major vertical edge. Shirdhonkar et. al. [6] had identified eight important features as height, height to width ratio, pixel density, location of the document, maximum and average horizontal run length (number of continues pixels in a row), horizontal transaction and vertical transaction. Thus, both researchers had suggested pixel density as an important feature.

For the current study, 6 features were identified in order to cluster the words successfully. Those are height, pixel density, pixel distribution, vertical projection variance, major vertical edge and major horizontal projection difference.

1) Height

The height of each word was taken into account. Generally, the height of handwritten words is higher than the height of printed words.

2) Pixel density

The image of words contains only black and white colour pixels. In white background, the characters are written in black colour. This feature calculates the percentage of black pixels in given area. This feature was chosen because, the printed words used to have high density compared to handwritten words as printed words are more compact than hand written words.

3) Pixel distribution

Normally, the questions, printed characters, in a questionnaire are given in both upper care and lower case. According to this system, the answers, handwritten characters, are given in upper case only. Therefore, if the image divides into two parts, lower part and upper part, the difference of pixel densities of two parts in printed words are higher than the difference of densities of hand written words. Therefore, the absolute value of the difference of upper part density and the lower part density had taken to the account.

4) Vertical projection variance

This is analyzing the pixel distribution for each column. The number of black pixels of each column was calculated and stored in an array. Afterwards the variance of that array was taken to the account. Normally, the variance for printed words is higher than the variance for handwritten words. Example was given in Fig. 4.

5) Major vertical edge

Using the array stored the number of black pixels of each column; the maximum value could be obtained. Generally, hand written characters contain curved lines instead of straight lines. Therefore, this value makes significant in printed words.

6) Major horizontal projection difference

Analyzing project profile brings much information. The number of black pixels in each row was stored in an array. And the difference of two near values is calculated. The maximum different value was taken to the account. Consider Fig. 5. When using both upper and lower case characters, the difference in the changing point of β takes a large value. But it is not that different in using only upper case

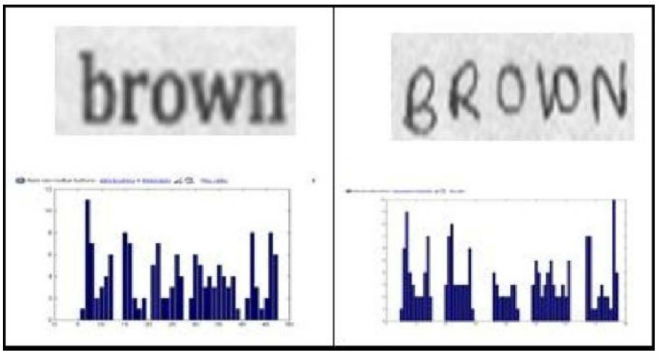

Fig. 4. Vertical projection variance

.

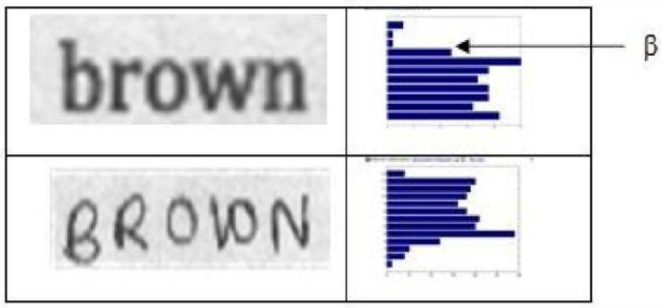

Fig. 5. Major horizontal projection difference

D. Feature selection for classification

Feature extraction is the best approach to identify each character as two characters will never be overlap with each other even if they are the same. Number of vertical and horizontal lines cannot be used as features, since the handwritten characters usually include curves instead of lines. Thus, the image was divided into 16 sub matrices and the pixel density was calculated for each matrix. Mamedov et. al. [7] used feed forward neural network to identify characters. The drawback of this method is time consuming. Emary et. al. [8] illustrates that Probabilistic Neural Network (PNN) often learns quickly than many neural network models. Thus, PNN was trained in order to identify characters.

III. Results and Discussion

This study was done with the aim of introducing a system which can automate the data entering process of a survey. Six features were identified in order to segment the words into two groups: height, pixel density, pixel distribution, vertical projection variance, major vertical edge and major horizontal projection difference.

The Agglomerative hierarchical method [9] was used to cluster the words. It is a bottom-up strategy which starts by considering each object as its own cluster and merge these clusters using the distance among clusters. Thus, there are many distance functions and linkage functions.

When calculating a distance, there are several requirements which need to satisfy [9]. A distance should be a nonnegative number. The distance of an object to itself should be 0. The distance function should be symmetric. Finally, the distance from object i to object j should be no more than making a detour over any other object h (triangular inequality). Both Euclidean distance, given in equation 1, and Manhattan distance, given in equation 2, satisfy these requirements.

$$d_{i,j} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2} \quad (1)$$

$$d_{i,j} = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{in} - x_{jn}| \quad (2)$$

A distance should be linked using proper linkage function. There are several linkage functions. If the clustering process is terminated when the distance between nearest clusters exceeds the arbitrary threshold, it is called a single linkage algorithm. Complete linkage distance is the

algorithm where the clustering process will be terminated when the maximum distance between nearest clusters exceeds an arbitrary threshold [9].

In order to choose the best linkage function and the distance function, the collected words using the questionnaire was clustered using each linkage function with distance function. Results shows that the clustering provides better results when using Euclidean distances and complete distance linkage as it maximize the Cophenetic correlation coefficient as 0.8874.

These features are sufficient to obtain a recall accuracy of 98%. Fig. 6 shows the Dendrogram plot drawn for this hierarchical clustering. Table 2 shows the misclassification rate for each feature. According to the results of table 2, it shows that pixel density was more significant than other features. Therefore, pixel density was chosen to name the clusters. For an example, if the pixel density of cluster 1 is greater than pixel density of cluster 2, it is clear that the hand written words were clustered as cluster 2 as hand written words do carry a low pixel density. Thus, this mechanism was used to identify clusters.

In order to identify the characters of handwritten words, the character image was pre processed and 16 partial densities are found as features. Therefore, an image contains 16 variables as input values. It was found that Probabilistic Neural network (PNN) was more accurate than feed-forward neural network as it calculates the global error where feed-forward neural network finds local error. Specht [10] suggest that PNN can be use for real time data.

TABLE II. THE EFFECT OF FEATURES

Feature	Accuracy
Height	72.54%
Pixel density	94.039%
Pixel distribution	86.275%
Vertical Projection Variance	90.117%
Major Vertical edge	78.43%
Major horizontal projection difference	82.35%

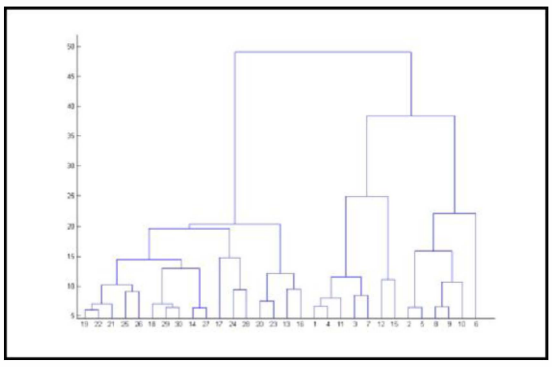

Fig. 6. A Dendrogram plot for Agglomerative hierarchical clustering

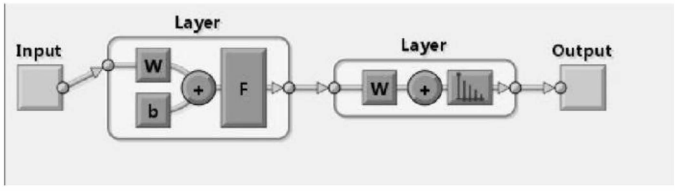

Fig. 7. Architecture of PNN

Beale et at. [11] describes PNN which was available in MATLAB. As given in Fig. 7, PNN use 3 layers in the classification process. In 1st layer, a new vector Z will be generated as given in equation 3 where X is the input matrix and W is the weight matrix [12]. Then the Euclidean distance was calculated for the Z matrix in order to create the distance matrix [11]. These elements were multiplied, element by element by bias and sent to the radial basis transfer function.
This function calculates layer's output from its net input and returns a matrix A, of the redial basis function applied to each element [11].

In second layer, it use Competitive transfer function to transfer the matrix A where in a given column, assigning 1 for the maximum value and 0 for rest.

$$Z=X.W \quad (3)$$

Donald et al.[10] shows that PNN can use for high dimensional data. PNN follows Gaussian distribution. Thus, assuming that the large amount of data behaves according to Gaussian distribution, the spread of dataset should be calculated initially. This was done by trial and error method and found that the maximum accuracy obtain when the spread is 14.The recall accuracy obtain using this method is 71.4% thus the system had difficulties of separating the number 0, letter O, number 1, letter I, number 2, letter Z and number 5, letter S.

IV. CONCLUSIONS

The features to cluster the words into two groups can be figure out as height, pixel density, pixel distribution, vertical projection variance, major vertical edge and major horizontal projection difference. It was found that height is the most important feature in order to cluster printed words and hand written words among other features. In order to recognize characters, the image was divided into 16 sum matrix and the pixel density of each sub matrix was calculated in order to recognize the character. Accuracy rate for clustering handwritten and printed words is 98% and for character recognition is 71.4%

Main 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 was a reason to reduce the accuracy of recognizing characters. Still, the system provides a better solution to automate the data entering of a questionnaire by providing high efficiency.

A. Findings

- The features to cluster the words into two groups can be figure out as height, pixel density, pixel distribution, vertical projection variance, major vertical edge and major horizontal projection difference.
- Height is the most important feature in order to cluster printed words and hand written words among other features
- In order to recognize characters, the image matrix is not a good approach to use as inputs.
- Therefore the image was divided into 16 sub matrix and the pixel density of each sub matrix was calculated in order to recognize the character

REFERENCES

[1] Tegang, S. P., Emukule, G., Wambugu, S., Kabore, I., & Mwarogo, P., (2009), A comparison of paper-based questionnaires with PDA for behavioural surveys in Africa, *Journal of Health Informatics in Development Countries,* **3**, 22-25

[2] Bartlett, J. E., Kotrlik, J. W. & Higgins, C.C (2011). Organizational Research:Determining Appropriate Sample Size in Survey Research, Information Technology, Learning and Performance Journal, 19(1).

[3] Kavallieratou, E., Dromazou, N., Fakotakis, N. and Kokkinakis G., (2003), An Integrated system for handwritten document image processing, *International Journal of Pattern Recognition and Artificial Intelligence*, **17**, 617-636

[4] Herath & Medagoda, (2006),Preprocessing Engine of the Optical Character Recognition System for Sinhala Scripts, pan localization project, 11-13

[5] da Silva, L.F. Conci, A. and Sanchez, A. (2009), Automatic Discrimination between Printed and Handwritten Text in Documents, Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on, 261 – 267

[6] Shirdhonkar M.S, Manesh B. Kokare, (2010), Discrimination between Printed and Handwritten Text in Documents, *International Journal of Computer Applications*, 131

[7] Mamedov, F. and Hasna, J. F. A., (2006), Character recognition using Neural Networks, Proceedings of the 2006 International Conference on Artificial Intelligence, 728-733

[8] Emary,I.M.M.E.l and Ramakrishnan, S., (2008), On the application of various Probabilistic Neural Networks in Solving Different Pattern Classification Problems. *World Applied Sciences Journal*, **4**, 772-780

[9] Data mining concepts and techniques

[10] Donald F. Specht, Probabilistic Neural Networks for classification, Mapping or associative memories.

[11] M. Beale, H. Demuth, " Neural network toolbox," For Use with MATLAB, User\'s Guide, The Math Works, Natick, 1998

[12] Online EMG Signal analysis for Diagnosis of Neuromuscular diseases by using PCA and PNN