

Short answer question examination using an automatic off-line handwriting recognition system and the Modified Direction Feature

Hemmaphan Suwanwiwat and Michael Blumenstein

School of Information and Communication Technology, Griffith University
Queensland, Australia

hemmaphan.suwanwiwat@griffithuni.edu.au, M.Blumenstein@griffith.edu.au

Abstract—Handwriting Recognition is one of the most intensive areas of study in the field of pattern recognition. The automatic assessment of exam scripts [1] can benefit from off-line handwriting analysis methods. Off-line automatic assessment systems can be an aid for teachers in the marking process; it reduces the time consumed by the human marker. Recently developed systems in this area have achieved an encouraging assessment yield range executed under constrained conditions. There has not been any recent work in the development of off-line automatic assessment systems using handwriting recognition, even though such systems will clearly benefit the education sector. There is a significant relationship between feature extraction and the recognition response yield; therefore the current research proposes the use of the Modified Direction Feature (MDF) extraction technique [2] and neural networks to develop an automatic assessment system for marking short answer questions. The system has high assessment accuracy (up to 92.31% for hand printed and 91.67% for cursive handwritten). The proposed system also includes marking criteria to augment its accuracy.

Keywords—Off-line handwriting recognition; Modified Feature Extraction; Backpropagation; Artificial Neural Networks; Automatic Assessment System

I. INTRODUCTION

Handwriting Recognition consists of on-line and off-line handwriting recognition techniques. An on-line system recognises handwriting in real time, which means analysis, occurs whilst the text is being written [3]. Devices such as a digitiser or an instrumented stylus are used to capture the written information whilst performing the writing [4]. On-line handwriting recognition can be seen in applications such as smart phone devices [5] as well as in automatic transcription of multilingual documents [6]. Off-line handwriting recognition on the other hand, recognises the written document. By using a scanner, the hard copy document is commonly transformed into a binary pattern [3] allowing the recognition system to process the binarised handwritten image. Some well known applications [4] that benefit from off-line handwriting techniques include Signature verification, Postal Address Interpretation, and Cheque Verification. These applications are practical, and in use worldwide.

Off-line handwriting recognition has certain disadvantages compared with on-line handwriting recognition systems. One main disadvantage is that the on-line technique captures the real-time information of the writing which is important in the recognition process,

including order of the strokes, stroke direction, and the speed of the writing within each stroke. Conversely, off-line handwriting recognition techniques require the image to be binarised and preprocessed to extract important features prior to the recognition process. The above-mentioned processes are expensive and important information could be lost when these steps are performed [3].

There have been recent advances in off-line handwriting recognition, such as feature extraction techniques, which could be used in automatic assessment systems in order to increase recognition rates. Hence the proposed research uses the Modified Direction Feature (MDF) extraction technique [2] for short answer question examination using an automatic off-line handwriting recognition system (AOHRS) for marking short answer questions with the objective of gaining greater assessment yields with limited or no constraints.

Automatic assessment is one of the applications of off-line handwriting recognition. Assessment of handwritten examinations is a difficult task; it requires the marker's concentration, precision and it is time consuming. Having an automatic handwritten examination assessment system would obviously be advantageous for markers to overcome some of the critical processes. Aside from improving accuracy, the aim of this research is also to improve the automatic assessment system usability. The proposed system will also include marking criteria to augment its accuracy and usability. Partly correct answers will also be marked.

The remainder of this paper is divided into four sections. Section II describes the methodology employed for this research. Section III details the results obtained and Section IV puts forward a discussion and analysis. Finally, conclusions are drawn in Section V and future work is also described.

II. METHODOLOGY

This section discusses the methodology and techniques used in conducting the research. The topics in this section include the proposed system method block diagram, data collection, proposed methodology, and the experimental setup.

The proposed methodology includes data collection, image processing and utilising the MDF technique in conjunction with different classification techniques in order to achieve the best results.

The block diagram below illustrates the proposed system methodology and processes. The individual components of the system are discussed in the following sub-sections.

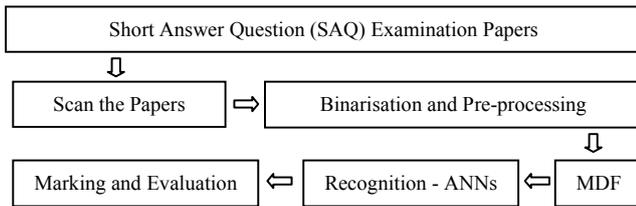


Figure 1. Block Diagram illustrating the proposed system

A. Data Collection

The proposed system is concerned primarily with assessing short answer questions from examination papers. The examination used to evaluate the proposed system is a short answer question assessment with no multiple choice or answers to choose from. The questions are concerned with fundamental Information Technology knowledge. The answers to the questions are close ended.

Examination papers containing 10 short answer questions were used for conducting experiments in order to gain 130 handwritten examination papers. The examination papers included 65 hand printed (HP), and 65 cursive handwritten (CH) samples from both male and female participants. 52 examination papers of each handwritten type were used as the training and testing datasets for the artificial neural networks (ANNs), and the remaining 13 of each type were used as the testing set for each AOHRs evaluation.

B. Datasets

For both HP and CH word samples to facilitate ANN training and testing, 80% of the available samples of each question were used as the training dataset, and 20% were used as the testing dataset. Numbers of each word in each dataset are varying. This is due to the number of available words obtained from the examination papers.

The datasets used in the experiments include HP, CH, and hand printed and cursive handwritten words combined (HPCHC). HP and CH datasets were sorted by the character cases. The HPCHC training dataset was sorted by their types and the character cases. The character cases in each training dataset include 3 cases which are uppercase, lowercase, and mixed-case.

The testing datasets were divided into 2 categories, which are HP and CH. The testing datasets contained all of the 3 cases.

C. Preprocessing

All examination papers were scanned and binarised. Salt and pepper noise removal and segmentation were performed. However, a slant correction process was not included. The images were only rotated into the best position possible on the base line. Most of the images still have noise and, they were used as they are.

D. Feature Extraction

The MDF technique was chosen due to its ability to extract important feature from images, so that the highest recognition rate could be achieved. The MDF feature vector

creation process is based on the calculation of direction and transition features from background to foreground pixels in the vertical and horizontal directions [2]. Originally the MDF was created to extract information from characters using direction transitions and transition feature information. Hence, the technique was originally developed to analyse information at the character level. However, the proposed system will be implementing the MDF to extract features from the whole word. Likewise, the recognition process will utilise whole word information rather than recognising the handwriting at the character level.

E. Classification

ANNs were trained with the resilient backpropagation algorithm and was selected above all others to address the problem of magnitudes of the partial derivative effects when using the sigmoid function.

F. Marking Criterion

A criterion for marking is important as it enables the system to mark the examination answers according to their quality. Human assessors generally consider the quality of the answers, and give partial marks to partially correct answers rather than mark the answers as zero or one. This also applies to AOHRs. The systems are clearly more usable and benefit the students being examined as they allow partially correct answers to be marked accordingly. The marking criterion was used in the marking phase. If the recognised word is a correct answer, the mark is given according to the marking scheme. For example, in the question "What does IT stand for?" the answer to this question contains 2 words which are "information" and "technology". If the answer contains any of those words, half marks will be awarded. The system's enhanced ability to mark more realistically partially resulted from the inclusion of a marking criterion into the system.

G. Experimental Settings

ANN settings are divided into 3 main categories, whereby there are 12 outputs, 13 outputs (the 13th output is a reject neuron), and 24 outputs. The 12 and 13 output ANNs were used for individual HP datasets and individual CH datasets. It was also used for HPCHC. The 24 output ANNs were only used to train and test the HPCHC.

The number of hidden units applied during ANN training was experimentally set from 3 up to 150 hidden units. The number of iterations set for training increased from 100 up to 18000. All ANNs were trained with a learning rate of 0.1, and momentum rate of 0.1.

H. Classification Criteria

The results of the experiments were determined by 3 different criteria, which are:

The first criterion is the classification criterion that uses the threshold of 0.5 and above from the ANN's output to determine if the word will be recognised as its output. If the highest threshold output is equal to or greater than 0.5, it will

be recognised as that output word. If the output word matched with the correct output, it is then recognised as the correctly recognised word. If all of the output threshold values are lower than 0.5, the highest threshold output will be compared if it is an unsure or incorrect word. If the output with the highest threshold has a value below 0.5 and equal to or above 0.4, it will be classified as an unsure word and will need to be manually marked. Otherwise the output of that testing word will be classified as an incorrect word.

The second criterion is almost identical to the first one, except it determines if the word recognised is correct according to the highest threshold found, regardless if the value of the threshold is at least 0.5. This criterion also investigates the percentage of the correctly recognised output with a low threshold value. This percentage is obtained by comparing the number of low threshold correct outputs with the total number of correct outputs where the threshold is higher or equal to 0.5.

The third criterion is an additional one which was only applied to the neural networks having 24 outputs. The criterion allows the output to be recognised regardless of the type of word (hand printed or cursive).

I. AOHRs Evaluations

The best classifications evaluated by the testing datasets were selected to be tested again with the unseen testing datasets before they were applied to the AOHRs. The unseen testing datasets included 13 HP and 13 CH examination papers. Only the best classification of each word type (HP and CH) was applied to the AOHRs. The AOHRs were divided into 2 systems which are HPAOHRs and CHAOHRs.

III. RESULTS

This section presents results of both HP and CH classification and accuracy rates. Selected classifications which were attained through evaluation with the testing datasets (20% from 52 examination papers of each writing type) were tested again with the 13 unseen examination papers of each type. The HP classification rates from the unseen testing datasets are displayed in sub-section A, and the classification rates for CH text are displayed in sub-section B. Sub-sections C and D display the results of HP and CH word accuracy rates respectively.

A. Hand Printed Classification Rates

The best trained classifiers were tested with the 13 unseen HP examination papers. The top results when evaluated with the first (CR1), second (CR2), and third criteria (CR3) from each setting are shown in Table I.

TABLE I. HAND PRINTED CLASSIFICATION RATES EVALUATED WITH THE FIRST, SECOND, AND THIRD CRITERIA

Output No. / Training Dataset	CR1 %	CR2 %		CR3 %	Hidden Units	No. of Iterations
		Class. Rate %	Thres. < 0.5 %			
12 / HP	85.26	87.18	2.42	N/A	43	7250
12 / HP & CH	91.03	92.31	1.56	N/A	61	10000
13 / HP	84.62	85.90	2.54	N/A	29	6900
24 / HP & CH	N/A	N/A	N/A	82.69	101	13000

The most successful neural network is the 12 output neural network trained with the HPCHC dataset. This most successful neural network is used as the classifier for the HPAOHRs.

B. Cursive Handwritten Classification Rates

The best CH trained classifiers were also tested with the 13 unseen cursive examination papers. The best results when evaluated with CR1, CR2, and CR3 from each setting are shown in Table II below.

TABLE II. CURSIVE HANDWRITTEN CLASSIFICATION RATES EVALUATED WITH THE FIRST, SECOND, AND THIRD CRITERIA

Output No. / Training Dataset	CR1 %	CR2 %		CR3 %	Hidden Units	No. of Iterations
		Class. Rate %	Thres. < 0.5 %			
12 / CH	85.26	87.82	5.51	N/A	21	6900
12 / HP & CH	87.18	91.67	5.55	N/A	61	10000
13 / CH	83.97	87.82	4.88	N/A	48	5000
24 / HP & CH	N/A	N/A	N/A	76.28	101	13000

The most successful trained classifiers for CH happened to be the same as those for HP classification. That is the 12 output neural network with 61 hidden units that was trained with the HPCHC dataset at 10000 iterations. This most successful neural network was used as the classifier for the CHAOHRs.

C. Hand Printed AOHRs Evaluation

Marking scores obtained from using the HPAOHRs are compared to human manual marking scores. The accuracy rates are calculated by the marking score obtained from the system divided by the marking score from the human manual marking and subsequently multiplied by one hundred, or vice versa if the marking score from the system is higher than the manual marking score. An accuracy rate of 91.99% was achieved.

D. Cursive Handwritten AOHRs Evaluation

CHAOHRs was evaluated by using the same criterion as the HPAOHRs. An accuracy rate of 89.94% was achieved.

TABLE III. CLASSIFICATION AND ACCURACY RATE COMPARISONS BETWEEN HAND PRINTED AND CURSIVE HANDWRITTEN AOHRs DISCUSSION AND ANALYSIS

Examination Type	Classification Rate %	Accuracy Rate %
Hand Printed	92.31	91.99
Cursive handwritten	91.67	89.94

This section discusses the AOHRs's classification and accuracy rates. The classification rates were those attained from the amount of words recognised in total from testing the 26 unseen examination papers (13 of each handwriting type). Whereas, the accuracy rates attained are based on the analysis of the words recognised and the number of actual correct words in total. The classification and accuracy rates of both HP and CH writing are discussed in sub-section A. The comparison between the AOHRs and other automated assessment systems described in the literature can be found in sub-section B.

E. Classification and Accuracy Rates

The results in the previous section show that both HPAOHRs and CHAOHRs can achieve classification rates above 90% (92.31%, and 91.67% respectively). However, accuracy rates attained from both HPAOHRs and CHAOHRs (91.99% and 89.94% respectively) are lower than the recognition rates of their types.

After analysing the outputs from the classifications, it was found that the outputs can be divided into 4 categories, which are described below:

- 1) *True positive*: The outputs that are recognised, and are marked correctly.
- 2) *False positive*: The outputs that are recognised, and marked incorrectly.
- 3) *True negative*: The outputs that are rejected correctly.
- 4) *False negative*: The outputs that are rejected incorrectly. That is the correct outputs are classified as incorrect output or have not been recognised as correct output.

TABLE IV. PERCENTAGE OF THE OUTPUTS IN EACH OUTPUT CATEGORY

Output Category	HPAOHRs	CHAOHRs
	Percentage %	Percentage %
True Positive	80.77	78.85
False Positive	1.28	2.56
True Negative	12.18	12.18
False Negative	5.77	6.41

It was found that the true negative and the false positive outputs caused the accuracy rates to be lower than the classification rates. After analysing the testing datasets, some factors that would affect the recognition and the accuracy rates were found.

- 1) *Confusing words*: Answers from the examinees including some confusing words which were hard for the classifier to recognise. A couple of good examples are the

word input and output. The word input and output together were misrecognised 5 times when marking the CH examination paper, and 4 times when marking the HP examination paper. While the classifier was able to recognise the word "technique" as the true negative output, it could not classify the word "ROM" as the true negative output. This is not beyond understanding; the feature of the word ROM and RAM could be very similar especially in CH words.

2) *Poor writing*: Some of the handwriting samples are hard to read, even to the human assessor. This matter has a direct effect on the recognition rate.

3) *Poor resolution of images*: the resolution of the scanned examination papers for the AOHRs evaluation datasets was quite low. A lot of white noise was found. Some manual filling was performed as the image quality was quite poor.

4) *Poor positioning of words in word matrices*: The images used in all experiments have not been slant corrected. The images were used as they are, except with some images where rotation needed to be performed.

F. The Comparison between the AOHRs and Automated Assessment Systems in the Literature

The only research about off-line automated assessment systems found in the literature was proposed by Allan [1].

The existing automated assessment systems and the system proposed here are actually quite different. It is also hard to compare the classification and accuracy rates due to the differences between the nature of the systems and the criteria used to determine such rates. However the features of both systems will be discussed. The comparisons between the systems include the dataset, feature extraction, the system itself, assessment type, and recognition and accuracy rate. The discussion is mainly based on an existing system whereby its experiments were conducted in 2001 [7]. Some features of the other existing systems are also discussed here.

1) *Datasets*: The datasets used in the literature were quite varied. These included handwriting from children to adults, depending on the experiments conducted. The dataset from the literature which is discussed here was collected from 50 first year computing students. The examination paper was made up of eight multiple choice questions (MCQ). Each question has 3 responses to choose from. The examinees were required to write the answer from the 3 responses available.

The datasets used in this study included 130 examination papers, which was larger than the previous dataset. The examination papers were divided into 2 categories which are HP and CH. The examination was a short answer handwritten examination which contains 10 questions.

2) *Feature Extraction*: The existing system used a Structure Feature Extraction technique with a holistic HVBC recogniser [8]. The holistic recogniser works by recognising the shape of the word from features extracted from the whole word image. The Cursive Script Recognition technique was also used in the other experiment.

The proposed system used The MDF technique and builds upon Direction Feature Extraction [9]. In this study, MDF extracts features from the whole word.

3) *The System*: The existing system is a lexicon based system. The proposed system has no lexicon.

4) *Assessment*: The assessment reported in previous research was performed on MCQs. Each response consisted of a selection of three words. These words were used so that bridges between each word were created. The number of bridges were counted to consider the recognition responses which are valid, possible and invalid. For example, a valid bridge can be found when two words are correctly recognised in the correct order. The answers used in the existing system were restricted to the words provided. These results obtaining higher recognition rates are due to the responses and the nature of restricted answers from the examinees that are known prior to the assessment process.

The proposed system does not have restrictions or any constraints in the question answering process. Even though the questions are closed ended, the examinees are allowed to write the answer freely. The proposed system also has the ability to mark the examination paper based on the quality of the answer. The marking criterion enables the system to be used in a real world situation, where marking the answers is mainly based on their quality.

5) *Recognition and Accuracy Rate*: The existing system [1] was reported to be able to achieve the assessment yield range from 54% with 99% accuracy to 44% with 100% accuracy depending on the constraints, such as lexicon and bridges between the lexicons, and the response history applied. The main system which is discussed here correctly scored 54% of all responses with an accuracy rate of 99%. The rates were attained by using bridges between each word, which may be considered a constraint.

The proposed system using HP words achieved a recognition rate of 92.31% with 91.99% accuracy rate when used to assess the examination. The CH word achieved the recognition rate of 91.67% with 89.94% accuracy when used to assess the examinations. The recognition and accuracy rates were obtained without applying any constraints.

IV. CONCLUSIONS AND FUTURE WORK

The motivation of this research was to create an AOHRs that can be used in the real world. That is an accurate marking system which does not only have the ability to mark a multiple choice examination, but also a handwritten one. The system also needs to achieve the highest recognition rate with the highest accuracy possible, as the system is being developed for off-line automatic assessment.

In this study, both of the off-line HPAOHRs and CHAOHRs used the same classifier, therefore they were combined into one AOHRs at the end of the experiment. The system was able to recognise and mark both HP and CH examination papers successfully

The recognition rate of 92.31% with 91.99% accuracy rate was achieved when assessing HP examination papers. The CH examination paper assessment achieved a recognition rate of 91.67% with 89.94% accuracy. Applying marking criteria to the system was practical; the answers were evaluated by their quality. The results obtained by the proposed system were quite encouraging. It is hoped that from the results obtained, future research would allow for improvements to be made to the system.

From the research undertaken, three main improvements may be suggested to enhance recognition rates of hand printed and cursive words as well as the marking accuracy.

Firstly automated preprocessing techniques must be applied. The datasets both for training and testing should be appropriately cleaned and slant corrected. Secondly, the different settings of ANNs and their structures may be changed. For example, instead of having one or two output neurons for each word type (hand printed or cursive handwritten), more output neurons may be applied for each character case as well. This may allow higher recognition to be achieved. Thirdly, cross validation could be performed, so that the training data is appropriately verified during the training process. Enhanced training processes and alternate classifiers could result in high recognition and accuracy rates.

REFERENCES

- [1] J. Allan, *Automated Assessment Of Handwritten Scripts*, Doctoral Dissertation, Nottingham Trent University, 2004.
- [2] M. Blumenstein, X. Y. Liu, and B. Verma, "A modified direction feature for cursive character recognition." *IEEE International Joint Conference on Neural Networks*, vol. 4, pp. 2983-2987, July 2004, doi: 10.1109/IJCNN.2004.1381140
- [3] C.C. Tappert, C.Y. Suen, T. Wakahara, "The State of the Art in Online Handwriting Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, August 1990, pp. 787-808, doi: 10.1109/34.57669
- [4] R. Plamondon, and S.N. Srihari, "Online and off-line handwriting recognition: a comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Jan 2000, vol.22, no.1, pp.63-84, doi: 10.1109/34.824821
- [5] E. Anquetil, and H. Bouchereau, "Integration of an on-line handwriting recognition system in a smart phone device," *16th International Conference Preceeding on Pattern Recognition*, 2002, vol.3, pp. 192- 195 vol.3, doi: 10.1109/ICPR.2002.1047827
- [6] A.M. Nambodiri, A.K. Jain, "Online handwritten script recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, January 2004, vol.26, no.1, pp. 124-130, doi: 10.1109/TPAMI.2004.1261096
- [7] J. Allan, T. Allen, N. Sherkat, P. Halstead, "Automated assessment: it's assessment Jim but not as we know it," *Sixth International Conference on Document Analysis and Recognition*, 2001, vol., no., pp.926-930, doi: 10.1109/ICDAR.2001.953921
- [8] N. Sherkat, R.J. Whitrow, R.G. Evans, "Wholistic recognition of handwriting using structural features," *IEE Colloquium on Document Image Processing and Multimedia (Ref. No. 1999/041)*, 1999, vol., no., pp.12/1-12/4M.
- [9] Blumenstein, B. Verma, H. Basli, "A novel feature extraction technique for the recognition of segmented handwritten characters," *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, 3-6 August 2003, vol., no., pp. 137- 141 vol.1, doi: 10.1109/ICDAR.2003.1227647