An Investigation of Novel Combined Features for a Handwritten Short Answer Assessment System

Hemmaphan Suwanwiwat
College of Business, Law and Governance (Information Technology), James Cook University, Australia
hemmpahan.suwanwiwat@jcu.edu.au

Umapada Pal
Computer Vision and Pattern Recognition Unit
Indian Institute, India
Umapada@isical.ac.in

Michael Blumenstein
School of Software
University of Technology Sydney
Australia
Michael.Blumenstein@uts.edu.au

***Abstract*—This paper proposes an off-line automatic assessment system utilising novel combined feature extraction techniques. The proposed feature extraction techniques are 1) the proposed Water Reservoir, Loop, Modified Direction and Gaussian Grid Feature (WRL_MDGGF), 2) the proposed Gravity, Water Reservoir, Loop, Modified Direction and Gaussian Grid Feature (G_WRL_MDGGF). The proposed feature extraction techniques together with their original features and other combined feature extraction techniques were employed in an investigation of the efficiency of feature extraction techniques on an automatic off-line short answer assessment system. The proposed system utilised two classifiers namely, artificial neural networks and Support Vector Machines (SVMs), two type of datasets and two different thresholds in this investigation. Promising recognition rates of 94.85% and 94.88% were obtained when the proposed WRL_MDGGF and G_WRL_MDGGF were employed, respectively, using SVMs.**

Keywords-off-line automatic assessment system; off-line handwriting recognition; Gaussian grid feature; modified direction feature; water reservoir feature

I. INTRODUCTION

There is only a small amount of research regarding off-line automatic assessment systems found in the literature even though paper-based examinations are still practically used world-wide. In larger classes, marking examination papers can be a difficult, prolonged, tiring, and error prone task. As a result a successful off-line automatic assessment system could be utilised to assist in marking so that the errors of marking (human errors) may be reduced.

The answers to the questions of the proposed off-line Short Answer question automatic Assessment System (SAAS) may contain several words; as a result partly correct answers will be marked in order to augment the system's accuracy and usability. This therefore reflects the practical assessment system usage.

Whole word recognition approach was employed in this study. The proposed SAAS employed six feature extraction techniques. These six techniques include two newly proposed techniques, two original, and two enhanced. The amendment to the original techniques was performed by integrating the centre of gravity feature (referred to as G – gravity) with the originals' feature vectors.

The two classifiers selected to be employed in this research were artificial neural networks and support vector machines. Two threshold values were employed in the proposed SAAS for the feature extraction techniques investigation. The datasets contained 3,000 and 3,400 samples from a total number of 100 writers. It was observed that many handwritten samples' legibility was reduced. This may be due to the stress the writers experienced and the fact that they were rushed while answering exam questions.

The remainder of this paper is organised as follows. Research methodology utilised in this research can be found in Section II, while the results attained and the discussion are described in Section III. Conclusions and discussion of the future research can be found in Section IV.

II. METHODOLOGY

The methodology and techniques employed in this investigation on the proposed SAAS were included in Fig. 1 block diagram. The proposed methodology, techniques and processes include short answer collection, image acquisition and preprocessing, the newly proposed and other feature extraction techniques (WRL_MDGGF, G_WRL_MDGGF, MDF, GGF, G_MDF, and G_GGF) utilisation.

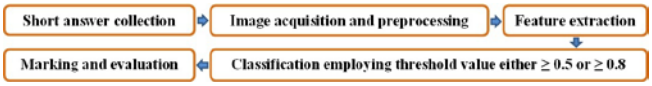

Figure 1. A block diagram of the research methodology and processes

These techniques were used in conjunction with artificial neural networks and support vector machines as the classifiers. The two classifiers as well as two threshold values of 0.5 and 0.8, were employed in order to investigate the efficiency of the proposed combined techniques, together with their original techniques on SAAS. The details of each of the system's components are described in the following sub-sections.

A. Short Answer Collection

The answers to the questions designed for the proposed SAAS were short. They contained a few words per question since they have been intentionally developed for assessing short answer questions from examination papers. The questions were closed. As a result, there was only one correct answer to each question; e.g. "What does CPU denote?" The

correct answer to the question can only be "Central Processing Unit".

The number of examination papers was set at one hundred; this number was designed to match larger class numbers of students. The samples, which were written with minimum constraints (no restriction on writing instruments) in the given writing space, were written by one hundred volunteers. Initially, there were 3,000 samples in the dataset (30 words × 100 writers).

As the samples were collected from examination papers, some incorrect answers were also included. Since 80% of the dataset was used for classifier training, additional data collection of each incorrect word was performed. The number of these additional answer words for each question was 8 to 42. Later, a further 400 handwritten samples of common incorrect answers were collected to be used in another training dataset. The total number of samples was increased to 3,400 samples.

B. Datasets

The experiments in this research employed 80% of the dataset for training and the remaining 20% for testing. There are two training datasets. The first training dataset (TR I) contained 2,400 samples; all samples in this dataset were correctly spelt and were correct answers to the questions. The second training dataset (TR II) contained 2,720 samples. This training dataset contained all of the correct answers to the questions as well as common wrong answers; all samples were correctly spelt. There were three testing datasets employed in this investigation namely, TE I, TE II, and TE III. TE I contained 600 samples; all the samples were correctly spelt and were correct answers to the questions. TE II and TE III contained both correct and incorrect answers to the questions; TE II contained 600 samples while TE III contained 680 samples.

Three datasets utilised in this research (DTS I, DTS II and DTS III) were created from the aforementioned training and testing datasets. DSI I contained TR I training dataset and TE I testing dataset. DSI II also contained TR I training dataset, however, employed TE II dataset. The last dataset, DSI III, contained TR II training dataset and TE III testing dataset. In total DSI I and DSI II contained 3,000 samples, where 2,400 samples were used for training and 600 samples were used for testing. DSI III contained 3,400 samples, where 2,720 samples were used for training and 680 samples were used for testing.

C. Image Acquisition and Preprocessing

All images were scanned with 300 dpi resolution and stored in grey-level format. They were binarised then segmented into word level. Boundary extraction, noise removal, skew and slant normalisation, as well as upper and lower contour extraction were performed on each image.

D. Feature Extraction Techniques

Feature extraction is an important process as it extracts the meaningful information that needs to be applied in the recognition process. Employing an efficient feature extraction technique would improve the recognition and accuracy rates.

Newly proposed combined feature extraction techniques called WRL_MDGGF and G_WRL_MDGGF, together with the original MDF, the original GGF, and their combined techniques, namely G_MDF [3], G_GGF [3] were employed in the proposed SAAS. These techniques had different feature vector sizes. The newly proposed WRL_MDGGF vector size is 1,569 compared to the proposed G_MDGGF of 1,587. The MDF vector size is 121 while GGF vector size is 864, 139 for G_MDF, and 882 for the G_GGF.

Fundamental features, the proposed combined feature extraction techniques, and the other combined techniques which were implemented in this study are explained below.

1) Fundamental features employed in the combined feature extraction technique creations. There are five main features namely, water reservoir, loop, centre of gravity, MDF [1], and GGF [2] employed in creating the proposed combined feature extraction techniques. The five fundamental features are described as follows:

a) Water Reservoir Feature (WRF): This technique was used to locate WRs found in upper and lower contours of images (refer to Fig. 2). The WR feature vector size is 392 for both upper and lower contour images (196 × 2 = 392) [3].

b) Loop Feature (LF): Loops may be found in some English alphabets. To obtain this feature, images are first divided into 3 zones (baseline, middle, top zone). Loops are then located within each zone. The loop feature vector size is 192 [3].

c) Centre of Gravity Feature (referred to as G): The centre of gravity feature was extracted from nine images comprised of a full boundary image, two equal horizontal and two equal vertical windows, and four equal windows which were obtained from dividing each full image into four equal windows. The centre of gravity vector size is 18 [3].

d) The Modified Direction Feature (MDF): The MDF [1] was first created to extract information from characters using direction transitions and transition feature information. This study employed the MDF at word level rather than at character level. The proposed system implemented the MDF to extract features in a heuristic approach. The MDF vector size is 121.

e) The Gaussian Grid Feature (GGF): The GGF [2] employs pattern contours as its input. A Gaussian smoothing filter (σ =1.2) is applied to each directional 12x 12 matrix. The size of the feature vector is 864.

2) The proposed combined feature extraction techniques. There are two feature extraction techniques proposed in this research. The first combined technique is called Water Reservoir, Loop, Modified Direction and Gaussian Grid Feature, and the second combined technique is called Gravity, Water Reservoir, Loop, Modified Direction and Gaussian Grid Feature. The proposed

combined feature extraction techniques are described as follows.

a) The proposed Water Reservoir, Loop, Modified Direction and Gaussian Grid Feature (WRL_MDGGF): The newly proposed WRL_MDGGF was created based on four features being WRF, LF, MDF and GGF. These features were selected to create a combined feature extraction technique due to their ability to successfully extract important features from images. Furthermore, their properties can be found in some or all English alphabets, which have enabled accurate recognition rates to be attained in a number of applications [1], [2], and [3].

The WRL_MDGGF was employed in this research to extract features at the word level. It extracts features from both full stroke and full boundary contour images. The WRF and LF were extracted first, and then MDF feature and GGF feature extraction took place. The WRL_MDGGF vector size is 1,569 which was obtained from 392 of WRF + 192 of LF + 121 of the MDF + 864 of the GGF.

b) The proposed Gravity, Water Reservoir, Loop, Modified Direction and Gaussian Grid Feature (G_WRL_MDGGF): The newly proposed G_WRL_MDGGF was created in an attempt to increase the recognition and accuracy rates. The reason for adding centre of gravity feature into the WRL_MDGGF vector was because positive outcomes were obtained when enhancing the original MDF, and GGF with centre of gravity feature [3]. The G_WRL_MDGGF feature vector is 1,587 which was obtained from 392 of WRF + 192 of LF + 121 of the MDF + 864 of the GGF 18 of the centre of gravity feature.

3) Other combined feature extraction techniques employed in this research. The two additional feature extraction techniques employed in this research were 1) Gravity, Modify Direction Feature (G_MDF), and 2) Gravity, Gaussian Grid Feature (G_GGF). Details of each technique are described below.

a) The Gravity, Modified Direction Feature (G_MDF): The recently proposed G_MDF [3] is a combined technique combining centre of gravity feature into its feature vector. The G_MDF vector size was 139; this obtained from 121 of the MDF + 18 from G vector).

b) The Gravity, Gaussian Grid Feature (G_GGF): The recently proposed G_GGF [3] is a combined technique adding centre of gravity feature into its feature vector. The G_GGF vector size was 882 (864 from the GGF + 18 from G vector).

E. Classification and Experimental Settings

There were two classifiers employed in conducting this research, namely Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs). With ANNs, the resilient backpropagation algorithm was utilised. With all experiments, ANNs were trained utilising either 2,400 or 2,720 samples, then tested with either 600 or 680 samples depending on the datasets (DSI I, II, or III – refer to sub-section II-B). The number of hidden units investigated was experimentally set from 50 up to 130 hidden units, incrementing by 1 at a time. The number of iterations set for training increased from 50 up to 1000, incrementing by 50 at a time.

The radial basis function SVMs with four-fold cross validation was used across all 3,000 and 3,400 handwritten samples. A four-fold cross validation was performed to get consistent and meaningful results; C parameter of the SVM was set at 30.

In all experimental settings and structure, there were 30 or 34 outputs for the 30 or 34 answers (words). All handwritten types of the same word belonged to the same output. For example “folder”, “FOLDER”, and “FoLDer” were classified as the same output.

F. Assessment Criterion

Assessment criterion is an important aspect of the SAAS as it enables the system to mark the examination answers according to their quality. This system marking scheme is clearly more usable and will benefit the students being examined, as it allows partially correct answers to be marked accordingly.

The assessment criterion was used in the marking phase. If the recognised word was a correct answer, the mark was given according to the marking scheme. For example, in the question "What does CPU denote?" the answer to this question contains 3 words which are “Central”, “Processing”, and “Unit”. If the answer contained any of those words, partial marks were awarded.

G. Classification Criteria

There were two threshold values, which were ≥ 0.5 and ≥ 0.8, utilised as classification criteria. The classification criteria were used to ensure that the recognised words passed through the marking phase with confidence (not just any recognised words with any threshold values).

For both threshold values, once each of the highest threshold valued outputs had been obtained, it was checked to evaluate if its value was more than or equal to 0.5 or 0.8 (depending on which threshold value was used). If it was, then it was recognised as the output and eligible for the marking process.

If the output threshold value was lower than 0.5 or 0.8 (depending on which threshold value was used), it would be classified as an ambiguous word and would need to be manually marked.

H. SAAS Evaluations

The SAAS evaluations employed three rates; these rates were 1) correctly recognised rate 2) assessment accuracy rate, and 3) efficiency rate. The correctly recognised rate was evaluated by using DSI I dataset. However, correctly recognised rate, assessment accuracy rate, and efficiency rate were evaluated by employing DSI II and III. Only the best recognition outcomes using each feature extraction technique were applied individually to the proposed SAAS. Correctly Recognised Rate (CRR) was calculated by using the total number of words that were recognised correctly (C) and the Total number of words in the testing dataset (T) that is:

$$CRR = (C / T) \times 100$$

The assessment accuracy rate is the rate which indicates the accuracy of the proposed system when the recognised words matched the answers to each of the questions, whilst rejecting words classified as being ambiguous for manual assessment. The mis-recognised rates were not included.

Assessment Accuracy Rates (AAR) were obtained by summing up the CRR and the rejection for Manual assessment rate (M) that is: $AAR = ((C + M) / T) \times 100$

The Efficiency Rate (ER) of the SAAS is the product of the CRR and AAC, and it was calculated by:

$$ER = (CRR \times AAC) / 100.$$

III. Experimental Results and Discussion

This section presents results including recognition rates obtained from the ANNs and SVMs classifiers, trained and tested using each Feature Extraction Technique (FET). Selected classifications were attained through evaluation of the testing datasets. The best result from each feature extraction technique was applied to the SAAS. The SAAS Correctly Recognised Rate (CRR), the Assessment Accuracy Rate (AAR) and the Efficiency Rate (ER) are discussed here (see Section II-H for these rates' descriptions).

1) Recognition Rates Attained from Employing DSI I dataset: For DSI I, both training and testing datasets contained only correctly-spelt correct answers to the questions. Total number of samples in this dataset was 3,000 (see Section II-B). The best recognition rates, obtained when employing each feature extraction technique in conjunction with ANNs and SVMs, can be found in TABLE I.

TABLE I. Recognition Rate (RR) Comparisons of Different Feature Extraction Techniques Employed, using DSI I Dataset

DSI I – 3,000 samples		
Feature Extraction Technique	***ANNs RR (%)***	***SVMs RR (%)***
MDF	97.33%	96.13%
GGF	96.17%	95.63%
WRL_MDGGF	93.17%	**98.20%**
G_MDF	**97.67%**	96.43%
G_GGF	97.00%	96.37%
G_ WRL_MDGGF	94.00%	**98.30%**

From Table I, it can be seen that the best RR of 98.30% (compared to 96.13% and 95.63% of the original MDF and GGF, respectively) was attained by employing the proposed G_WRL_MDGGF, using SVMs as the classifier. When ANNs were utilised as the classifier, the combined G_MDF technique was able to achieve the best RR of 97.67%.

2) Recognition Rates Attained from Employing DSI II dataset: The DSI II dataset contained 3,000 samples. While the TR I training dataset, which contained 2,400 samples, was used for training, the remaining 600 samples (TR II) were used for testing. As described in Section II-B, TR I contained only correctly spelt correct answers to the questions; however, TE II contained some incorrect as well as correct answers to the questions. The results of each of the FETs' best recognition rate are displayed in Table II.

TABLE II. Recognition Rate (RR) Comparisons of Different Feature Extraction Techniques Employed, using DSI II Dataset

DSI II – 3,000 samples		
Feature Extraction Technique	***ANNs RR (%)***	***SVMs RR (%)***
MDF	88.17	92.2
GGF	93.17	91.23
WRL_MDGGF	91.17	94.03
G_MDF	91	92.17
G_GGF	**93.83**	91.87
G_ WRL_MDGGF	92	**94.07**

It can be observed from Table II that the highest RR attained when DSI II was utilised was 94.07%. This rate was achieved when the proposed G_WRLGGF_MDF was employed, using the SVMs as the classifier. This rate was also higher than its original MDF and GGF counterparts' RR of 92.2% and 91.23%, respectively. With ANNs, the highest RR of 93.83% was obtained when the combined feature technique G_GGF [3] was employed.

TABLE III. Correctly Recognised Rate (CRR), Assessment Accuracy Rate (AAR), and Efficiency Rate (EF) Comparisons of Different Feature Extraction Techniques Employed using DSI II Dataset, Utilising Threshold ≥ 0.5

DSI II – 3,000 samples – employing threshold value ≥ 0.5			
Feature Extraction Technique	***CRR (%)***	***AAR (%)***	***ER (%)***
MDF	88.17	**99.17**	87.44
GGF	93.17	97.67	90.99
WRL_MDGGF	91.17	96.67	88.13
G_MDF	91	99	90.09
G_GGF	**93.83**	98.17	**92.11**
G_ WRL_MDGGF	92	97	89.24

There were two minimum threshold values employed in the experiments which were 0.5 and 0.8; details regarding these thresholds can be found in Section II-G. The experiment results attained when the threshold value ≥ 0.5 was employed are displayed in Table III, while the results obtained when the threshold value ≥ 0.8 was utilised are shown in Table IV.

TABLE IV. Correctly Recognised Rate (CRR), Assessment Accuracy Rate (AAR), and Efficiency Rate (EF) Comparisons of Different Feature Extraction Techniques Employed using DSI II Dataset, Utilising Threshold ≥ 0.8

DSI III - 3,000 samples – employing threshold value ≥ 0.8			
Feature Extraction Technique	***CRR (%)***	***AAR (%)***	***ER (%)***
MDF	59.17	99.33	58.77
GGF	79	99.17	78.34
WRL_MDGGF	**81.17**	97.67	**79.28**
G_MDF	68.67	**99.5**	68.33
G_GGF	78.83	99	78.04
G_ WRL_MDGGF	73.33	97.17	71.25

When the threshold value ≥ 0.5 criterion was employed (see Table III), it was found that the highest CRR of 93.83% with 98.17% AAR and 92.11% ER were attained when the combined G_GGF was employed. The highest AAR of 99.17% was achieved when the original GGF was employed as FET. However, since its CRR of 88.17% was rather low (lowest in the group), its ER of 87.44% was also reduced.

As can be seen from Table IV, when the threshold value ≥ 0.8 criterion was employed, it was found that the highest CRR of 81.17% with 97.67% AAR and 79.28% ER were

attained when the proposed WRL_MDGGF was employed. The highest AAR of 99.50% was however attained when the recently proposed G_MDF [3] was employed. Despite its high AAR, its ER was not high (68.33%). This was due to the fact that its CRR was not high and therefore its ER was reduced.

It can be noted that the proposed WRL_MDGGF was able to outperform its original MDF and GGF counterparts' correctly recognition rates (81.17% compared to 59.17 of MDF's and 79% of GGF's). The proposed WRL_MDGGF's efficiency rate of 79.28% was also the highest in the group which means it also outperformed the original MDF and GGF rates of 58.77% and 78.34%, respectively.

3) Recognition Rates Attained from Employing DSI III dataset: The DSI III dataset contained 3,400 samples. While the TR II training dataset, which contained 2,720 samples, was used for training, the remaining 680 samples (TR III) were used for testing. As described in Section II-B, both TR II and TE III contained incorrect as well as correct answers to the questions. The best recognition rates attained from the employed FETs are displayed in Table V. The experimental results, when the threshold values of ≥ 0.5 and ≥ 0.8, were employed are shown in Table VI and VII, respectively.

TABLE V. RECOGNITION RATE (RR) COMPARISONS OF DIFFERENT FEATURE EXTRACTION TECHNIQUES EMPLOYED, USING TRAINING DSI III

DSI III – 3,400 samples		
Feature Extraction Technique	***ANNs RR (%)***	***SVMs RR (%)***
MDF	92.06	92.79
GGF	**95.15**	92.41
WRL_MDGGF	92.94	94.85
G_MDF	90	92.88
G_GGF	94.26	92.97
G_ WRL_MDGGF	92.21	**94.88**

Upon observation the experiments results in Table V, it was found that the best recognition rate of 95.15% was obtained when the original GGF was employed on DSI III. However, when utilising the SVMs, the highest recognition rate of 94.88% was achieved when the proposed G_WRL_MDGGF was employed as FET. It was also observed that both of the proposed WRL_MDGGF and G_WRL_MDGGF RR of 94.88% and 94.85% outperformed their MDF and GGF counterparts' RRs of 92.79% and 92.41%, respectively.

When inspecting the results in TABLE VI, it was found that when threshold value ≥ 0.5 was utilised, the best CRR of 95.15% with AAR of 98.53% and ER of 93.75% was obtained when the original GGF was used as FET. The highest AAC of 99.71% was, however, achieved by employing the recently proposed combined features G_GGF [3] as FET.

In comparing the best AAR from employing the threshold value of 0.8 (see Table VII) rather than 0.5, it was found that the highest AAR was 99.85% which was attained when the G_MDF or MDF was employed. This rate was marginally higher than the best AAR of 99.71% obtained from employing the recently proposed G_MDF, utilising the threshold value of 0.5. This small improvement obtained from an expensive trade-off between CRR and AAR. It was also observed that the CRR of every FET employed was lowered, and as a result, their ERs were also reduced. The CRRs were reduced due to the fact that since the higher threshold value was enforced, fewer numbers of recognised outputs were allowed through marking phrase.

TABLE VI. CORRECTLY RECOGNISED RATE (CRR), ASSESSMENT ACCURACY RATE (AAR), AND EFFICIENCY RATE (EF) COMPARISONS OF DIFFERENT FEATURE EXTRACTION TECHNIQUES EMPLOYED USING DSI III DATASET, UTILISING THRESHOLD VALUE ≥ 0.5

DSI III – 3,400 samples – employing threshold value $\geq$ 0.5			
Feature Extraction Technique	***CRR (%)***	***AAR (%)***	***ER (%)***
MDF	92.06	98.97	91.11
GGF	**95.15**	98.53	**93.75**
WRL_MDGGF	92.94	97.06	90.21
G_MDF	90	**99.71**	89.74
G_GGF	94.26	99.12	93.43
G_ WRL_MDGGF	92.21	98.09	90.45

TABLE VII. CORRECTLY RECOGNISED RATE (CRR), ASSESSMENT ACCURACY RATE (AAR), AND EFFICIENCY RATE (EF) COMPARISONS OF DIFFERENT FEATURE EXTRACTION TECHNIQUES EMPLOYED USING DSI III DATASET, UTILISING THRESHOLD VALUE ≥ 0.8

DSI III – 3,400 samples – employing threshold value $\geq$ 0.8			
Feature Extraction Technique	***CRR (%)***	***AAR (%)***	***ER (%)***
MDF	64.71	**99.85**	64.61
GGF	82.06	98.82	81.09
WRL_MDGGF	76.03	96.91	73.68
G_MDF	66.91	**99.85**	66.81
G_GGF	**84.71**	98.09	**83.09**
G_ WRL_MDGGF	72.79	97.06	70.65

The comparisons between each FET's highest CRR, AAR, and ER are shown in Table VIII. It could be concluded that the highest AAR of 99.5% and 99.85% were achieved when DSI II and III were employed, respectively. Having trained them with DSI III, which contained common incorrect as well as all the correct answers to the questions, seemed to assist in improving the CRR, AAR, and ER. It was observed that by employing threshold value ≥ 0.8 rather than ≥ 0.5 to the classifiers, the AARs of some FETs were improved. However, as discussed earlier, even though the AARs were improved, the CRRs and ERs were decreased significantly. This was due to the fact that fewer outputs were allowed through the marking phrase.

TABLE VIII. THE HIGHEST CORRECTLY RECOGNISED RATE (CRR), ASSESSMENT ACCURACY RATE (AAR), AND EFFICIENCY RATE (EF) COMPARISONS OF DIFFERENT FEATURE EXTRACTION TECHNIQUES EMPLOYED USING DSI II AND III, UTILISING THRESHOLD ≥ 0.5 AND ≥ 0.8

Dataset / *T Value*	***FET***	***CRR (%)***	***AAC (%)***	***ER (%)***	***Best AAR From Highest CRR (%) taken from Tables III, IV, VI, and VII/ FET***
DSI II ≥ 0.5	G_GGF	93.83	98.17	92.11	99.17/MDF
DSI III ≥ 0.5	GGF	95.15	98.53	93.75	99.71/G_MDF
DSI II ≥ 0.8	WRL_MDGGF	81.17	97.67	79.28	99.5/G_MDF
DSI III ≥ 0.8	G_GGF	84.71	98.09	83.09	99.85/G_MDF,MDF

4) The Comparison between the proposed SAAS and other off-line word recognition techniques found in the literature: The comparison in this study was mainly performed with other off-line word recognition techniques found in the literature since there are not many studies sourced regarding SAAS.

TABLE IX. COMPARISONS BETWEEN DATASET SIZE (DZ), RECOGNITION RATE (RR) AND ACCURACY RATE (AR) OF THE PROPOSED SAAS AND OTHER SYSTEMS USING VARIOUS TECHNIQUES FOUND IN THE LITERATURE

System – Feature Extraction Technique	DZ	RR (%)	AR (%)
English Numeral Recognition – Hybrid Features (Moment of Inertia and Projection) [4]	3,500	91.7	91.7
English Character Recognition – Hybrid Features (Diagonal and Directional Based) [5]	5,200	95.96	N/A
Jawi Handwritten Recognition – Hybrid Features (Centre of Gravity, Zoning, pixel profiles, etc.) [6]	2,377	94.52	N/A
Handwriting recognition – Moment based features [7]	200	N/A	98.93
Persian Legal Amount Recognition – Zoning, Pixel Averaging , etc. [8]	4,500	80.88	N/A
– Children's Handwritten Responses – HVBC FET [9]	145	65.00	100
– Automated Assessment System - HVBC FET and Constraints employed [9]	1,077	54.00	99.00
Short Answer Automated Assessment System–G_GGF[3]	1,248	87.12	91.12
Proposed SAAS – Proposed WRL_MDGGF (SMVs)	3,400	94.85	97.06
– Proposed G_WRL_MDGGF(SVMs)		94.88	98.09
– The recently proposed G_MDF (SVMs) [3]		92.88	99.85

It can be observed that when comparing experiment results with other recognition systems [4]–[8], the proposed WRL_MDGGF and G_WRL_MDGGF, as well as the recently proposed G_GGF combined feature extraction techniques, were able to attain considerably high to comparable recognition and accuracy rates compared to the existing systems in Table IX. It could be noted however, that it was difficult to compare due to the dataset sizes and the nature of the words/characters/numerals utilised (e.g. bank cheque legal amount, different languages, and children vs. adult responses).

For SAAS comparisons (see Table IX), the proposed combined feature extraction techniques were able to achieve high recognition rates with comparable accuracy rates to those found in the literature. One of the automated assessment systems was able to obtain a RR of 54% with the AAR of 99%, while another system was able to achieve 65% RR with 100% AAR depending on the constraints employed [9]. The constraints which were used in the existing systems [9] were lexicon and bridges between the lexicons, and the response history. As a result, it should be noted that the AARs of 97.06% – 99.85% (attained by employing the newly/recently proposed feature extraction techniques) were achieved without applying any heavy constraints; and the proposed SAAS was able to attain the higher RRs of 92.88 – 94.88% compared to 54% and 65% of [9].

IV. CONCLUSIONS AND FUTURE WORK

The novelty of this research is the proposed combined feature extraction technique called the Water Reservoir, Loop, Modified Direction and Gaussian Grid Feature (WRL_MDGGF) and its enhanced technique called Gravity, Water Reservoir, Loop, Modified Direction and Gaussian Grid Feature (G_WRL_MDGGF). The proposed WRL_MDGGF and G_WRL_MDGGF were able to attain the highest recognition rates of 94.85% and 94.88%, respectively when employed DSI III dataset, utilising SVMs as the classifier. It was found that by employing a higher threshold value (≥ 0.8 rather than 0.5), the accuracy rates of some FETs were increased. These rates were increased from 0.16% to 0.88%. They were achieved with the trade-off between CRR and ER. In other words, since the higher threshold value was used, fewer outputs could pass (with high confidence) through to the marking phase. As a result, CRRs were reduced and caused the systems' efficiency (refer to ER) to be reduced.

Despite the costly trade-off between CRR, ER and AAR, it is very important to note that the improvements in AAR were important. Any increment in the assessment accuracy rate can be considered very important for SAAS as marking examinations incorrectly could cause a student to fail the exam. This study was able to achieve the best AAR of 99.85%.

Some suggestions for future work to improve the recognition rate, correctly recognised rate, assessment accuracy rate, and efficiency rate include the employment of alternative classifiers such as Hidden Markov Models (HMMs). Different algorithms and settings of ANNs and SVMs can be applied. Furthermore, hybrid classifiers (e.g. HMMs & SVMs, SVMs & ANNs) can be employed. Rather than utilising a whole word recognition approach, segmentation-based recognition may be applied to SAAS. More complex datasets (i.e. increasing from word to sentence level, larger dataset sizes, multilingual) can be collected and employed in future work.

REFERENCES

[1] M. Blumenstein, X. Y. Liu, and B. Verma. A modified direction feature for cursive character recognition. In International Joint Conference on Neural networks, volume 4, pages 2983-2987, 2004.

[2] V. M. Nguyen and M. Blumenstein. An application of the 2D Gaussian filter for enhancing feature extraction in off-line signature verification. In 11th International Conference on Document Analysis and Recognition (ICDAR 2011), pages 339-343, 2011.

[3] H. Suwanwiwat, V. Nguyen, M. Blumenstein, and U. Pal, A complete automatic short answer assessment system with student identification, In 13th International Conference on Document Analysis and Recognition (ICDAR 2015), pp.611-615, 2015.

[4] B. K. Prasad and G. Sanyal, A hybrid feature extraction scheme for Off-line English numeral recognition, 2014 International Conference Convergence of Technology (I2CT), pp. 1-5, 2014.

[5] J. Pradeep, E. Srinivasan and S. Himavathi, Performance analysis of hybrid feature extraction technique for recognizing English handwritten characters, 2012 World Congress on Information and Communication Technologies (WICT), pp. 373-377, 2012.

[6] A. Heryanto, M. F. Nasrudin and K. Omar, Offline Jawi handwritten recognizer using hybrid artificial neural networks and dynamic programming, 2008 International Symposium on Information Technology, pp. 1-6, 2008.

[7] P. Kumawat, A. Khatri and B. Nagaria, Comparative Analysis of Offline Handwriting Recognition Using Invariant Moments with HMM and Combined SVM-HMM Classifier, 2013 International Conference on Communication Systems and Network Technologies (CSNT), pp. 140-143, 2013.

[8] E. Shahriarpour and J. Sadri, Recognition of legal amount words on Persian bank checks using Hidden Markov Model, 2014 Iranian Conference on Intelligent Systems (ICIS), pp. 1-5, 2014.

[9] J. Allan. Automated Assessment Of Handwritten Scripts. PhD thesis, Nottingham Trent University, 2004.