An Investigation of Discrete Hidden Markov Models on Handwritten Short Answer Assessment System

Conference Paper · September 2018

CITATIONS
2

READS
50

5 authors, including:

Hemmaphan Suwanwiwat
James Cook University Brisbane
12 PUBLICATIONS **31** CITATIONS
SEE PROFILE

Abhijit Das
Indian Statistical Institute
48 PUBLICATIONS **410** CITATIONS
SEE PROFILE

Miguel A. Ferrer
Universidad de Las Palmas de Gran Canaria
293 PUBLICATIONS **3,338** CITATIONS
SEE PROFILE

Umapada Pal
Indian Statistical Institute
397 PUBLICATIONS **8,072** CITATIONS
SEE PROFILE

An Investigation of Discrete Hidden Markov Models on Handwritten Short Answer Assessment System

Hemmaphan Suwanwiwat[a], Abhijit Das[b,e], Miguel A. Ferrer[c], Umapada Pal[d], and Michael Blumenstein[e]

[a]Information Technology Academy, James Cook University, Cairns, Australia, art.suwanwiwat@jcu.edu.au

[b]Institute for Integrated and Intelligent Systems, Griffith University, Queensland, Australia, abhijit.das@griffithuni.edu.au

[c]IDeTIC, University of Las Palmas de Gran Canaria, Las Palmas, Spain, mferrer@dsc.ulpgc.es

[d]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India, umapada@isical.ac.in

[e]School of Software, University of Technology Sydney, Australia, michael.blumenstein@uts.edu.au

***Abstract*—This paper presents an investigation of an off-line automatic assessment system utilising discrete Hidden Markov Models. A set of geometric features were extracted from handwritten words and were later classified by HMMs. There were two training datasets employed in the experiments; the first training dataset contained all correct answers to the questions whereas another training dataset contained both correct and incorrect answers to the questions. Datasets contained 3,000 and 3,400 handwritten samples, respectively. The experiments yielded promising results whereby the highest recognition rate of 91.90% with a 100% accuracy was achieved on our database.**

Keywords—off-line automatic assessment system, Hidden Markov Models (HMMs), fixed-point arithmetic, geometric features

I. INTRODUCTION

Even though presently computer-based examinations have become widely accepted, paper-based examinations are still in use worldwide throughout all levels of education, including but not limited to secondary and tertiary levels. Despite the fact that paper-based examinations have been in use all these years, to the best of the authors' knowledge, literature regarding off-line automatic assessment systems is limited [1], [2] and [3]. Recognising off-line handwritten words is challenging when compared to recognising on-line handwritten words. There are a number of disadvantages in attempting to recognise off-line handwritten words because there is no real-time information available. Apart from that, whereas on-line recognition systems use both temporal and spatial information, only spatial information is available for off-line cases [4].

Recognising handwriting of students while answering questions in examinations can be considered difficult, as the students may be writing with significant stress and as a result could be writing in a way where legibility is reduced. Also there can be a high variance in the artefacts employed for examination (such answer sheet paper quality, colour, type of pen used, etc).

For essay or short-answer question assessment types, it is known that manually marking these types of exams is tedious, time consuming and most of all, error prone. To overcome this problem, an off-line Short Answer question automatic Assessment System (SAAS) is proposed in this paper. This study investigated the use of discrete Hidden Markov Models (HMMs) on short answer words.

HMMs have been used in both off- and on-line handwriting recognition systems. As stated by Plötz et al. [5], the sliding windows principle is an important milestone for successful Markov-model-based handwriting recognition, especially for off-line handwriting systems. HMMs are widely employed in automatic off-line recognition applications, including industrial ones. Hence, this study proposes a SAAS employing HMMs.

This study's contributions include:

1) Exploring the efficiency of employing discrete HMMs on the proposed short-answer question words. There have not been any experiments performed on the SAAS system using HMMs previously. This study shows encouraging results by employing the stated classifier.

2) Investigating the effect of the numbers of training samples on classification rates. The previous studies [3] employed 80% of the total number of samples in the datasets, whereas in this proposed research, the training datasets contained 10 – 50% and 80% of the total number of samples in the datasets. It was found that by employing HMMs, the best classification results were obtained when only 10 – 20% of the total proportion of the datasets were used to train the classifiers.

The features employed in the proposed system were geometrical features [6] which were based on two vectors that represent the envelope description and the interior stroke distribution in polar and Cartesian coordinates. Since HMMs were employed, no segmentation of the images was required.

There are two main training datasets utilised in this investigation. For the first type, the training dataset only contained correct answer handwritten samples to the questions. For the second type, however, the training dataset contained both correct and incorrect answer handwritten samples to the questions. There were also two main testing datasets, similar to the training datasets; the first testing dataset only contained correct answers to the questions, and the second one contained both correct and incorrect answers to the questions. There were altogether 18 sub-datasets employed in this study. In total, there were 3,000 and 3,400 handwritten samples in the two main

dataset types employed in this study. More details are described in Section II.

The remainder of this paper in divided into three sections as follows: Section II describes the research methodology employed in this study, while the report on experimental results can be found in Section III. Conclusions and discussion of the future research can be found in Section IV.

II. METHODOLOGY

Figure 1 illustrates the methodology employed in this investigation in relation to the use of discrete HMMs and the proposed SAAS. The proposed methodology, classification technique and processes including handwritten short answer words preparation, pre-processing, word segmentation, and HMMs are discussed in this section.

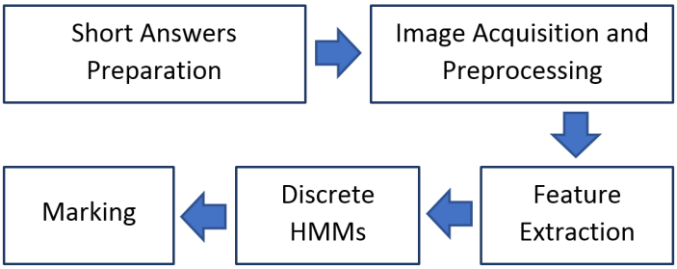


Fig. 1. A block diagram of the research methodology and processes

A. Short Answer Handwritten Words

The answers to the questions employed in this study were designed to be a few words per question, which suits the purpose of the proposed short answer question assessment system. The answers to the questions were straightforward for example "What does IT stand for?", The correct answer can only be "Information Technology" although the writers may write the words using different cases i.e. "information technology", "Information technology", etc.

The handwritten samples were obtained from datasets employed in [3]. There were 3,000 samples in the dataset which were obtained from 30 words written by 100 writers. A further 400 handwritten samples of commonly incorrect answers were collected to be used in another training dataset. The total number of samples was increased to 3,400 samples. Examples of handwritten short answers can be seen in Figure 2.

B. Datasets

There are two main types of training datasets and there are two main types of testing datasets. The first type of training dataset I (TR I) contained only correct answers to the questions. The second type of training dataset (TR II) contained both correct and incorrect answers to the questions. For testing datasets, the two main types of datasets were also applied. The first testing dataset (TE I) contained only correct answers to the questions whereas the second testing dataset (TE II) contained both correct and incorrect answers to the questions. From the four main datasets (TR I, TR II, TE I, and TE II), there are 3 training and testing dataset combinations.

TABLE I. NUMBER OF SAMPLES IN EACH COMBINATION SET

Dataset	No of Samples
CI	3,000
CII	3,400
CIII	3,400

The first combined dataset (CI) contained TR I and TE I which means that only correct answers to the questions were contained in this dataset. Since there are only correct answers to the question in this dataset, the size of the dataset is 3,000 samples. The second combined dataset (CII) contained TR I and TE II which means that the training dataset contained only correct answers to the questions whereas the testing dataset contained both correct and incorrect answers to the questions. There were 3,400 samples in this dataset since there are incorrect samples included in the dataset. For the last combined dataset (CIII), both correct and incorrect answers to the questions comprised both training and testing datasets. Same as CII, CIII contained 3,400 samples in its dataset (see Table I). The False Acceptance Rate (FAR) and False Rejection Rate (FRR) score statistics are also summarised in table

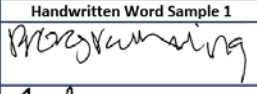
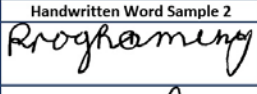
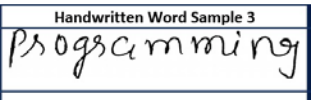
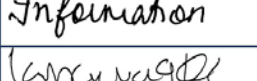
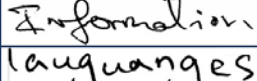
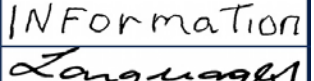
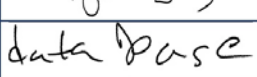
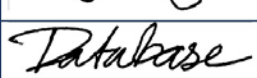
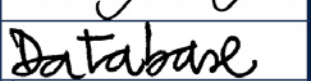
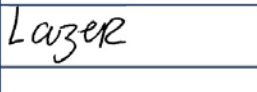
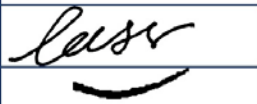
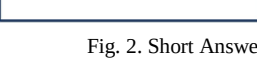
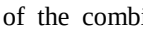
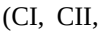
Word	Handwritten Word Sample 1	Handwritten Word Sample 2	Handwritten Word Sample 3
Programming			
Information			
Languages			
Database			
Laser			
Dash (students wrote this sometimes when they did not know the answer to the question)			

Fig. 2. Short Answer Samples

Each of the combined datasets (CI, CII, and CIII) were further divided into six sub-datasets. Each sub-dataset contained different amounts of training and testing samples. Table II below shows the percentages of amounts of samples in each of the sub-datasets. The numbers of training samples in training datasets were between 50 – 10% and 80%. Therefore, the numbers of

testing samples in the testing datasets were between 50 – 90% and 20% respectively (see Table II). The different numbers of training samples in the datasets were meant for the investigation on the efficiency of discrete HMMs towards the proposed SAAS.

TABLE II. EACH SUB-DATASET – NUMBER OF TRAINING AND TESTING PERCENTAGES

Sub-dataset	Training (%)	Testing (%)	FAR and FRR for CI	FAR and FRR for CII and CIII
S I	80%	20%	FAR=30*20 FRR=30*29*20	FAR=34*20 FRR=34*29*20
S II	50%	50%	FAR=30*50 FRR=30*29*50	FAR=34*50 FRR=34*29*50
S III	40%	60%	FAR=30*60 FRR=30*29*60	FAR=34*60 FRR=34*29*60
S IV	30%	70%	FAR=30*70 FRR=30*29*70	FAR=34*70 FRR=34*29*70
S V	20%	80%	FAR=30*80 FRR=30*29*80	FAR=34*80 FRR=34*29*80
S VI	10%	90%	FAR=30*90 FRR=30*29*90	FAR=34*90 FRR=34*29*90

C. Image Acquisition

All handwritten samples were scanned with 300 dpi resolution and stored in a grey-level format. The images were then binarised and segmented at the word level. Words were segmented and checked to ensure that there were no segmentation errors. Noise removal, skew and slant normalisations were performed on each image.

D. Feature Extraction Technique

Geometrical features [7] were employed in this investigation. Originally, this feature extraction technique was created for off-line signature verification. This investigation was also conducted to find out whether this technique is suitable for the verification of the students from their handwriting as well. The geometrical features are based on two vectors. They represent the interior stroke distribution of polar and Cartesian coordinates and the envelope description [7]. Outline detection and representation, feature vectors based on polar coordinates and feature vectors based on Cartesian coordinates are briefly described as follows:

1) Outline Detection and Representation: Morphological operations were used to calculate the outline. A dilatation was applied in order to reduce the word variability, after that the outline extraction process was simplified by a filling operation. After filling, a number of objects were detected, then a horizontal dilatation was performed until all the objects were connected. As a result, a sequence of the outline's Cartesian coordinates, being its length, was obtained.

1) Feature Vector Based on Polar Coordinates: In order to represent a handwritten word outline in polar coordinates, it was decided to select equidistant samples of the envelope and represent each sample as a three-component feature vector being 1) the derivative of the radius, 2) its angle, and 3) the number of black pixels that the radiuses crossed when sweeping from one selected point to the next.

The radius function is calculated as the number of pixels from the geometric centre to each outline selected point as:

$$d_1 = X_{t\cdot p} - C_x,\ d_2 = Y_{t\cdot p} - C_y$$

$$r_t = \max(d_1, d_2) + min(d_1,d_2)/4 \quad (1)$$

The angle of each selected contour sample is calculated by means of the arctan function implemented through a lookup table:

$$\theta_t = arctan(X_{nT/Tr}/Y_{nt/Tr}),\ t = 1, 2, \ldots, T_r. \quad (2)$$

2) Feature Vector Based on Cartesian Coordinates: This vector is also based on the envelope and the signature strokes density parameterisation, however in this scenario, using Cartesian coordinates. The envelope was divided through the geometric centre into top and bottom halves. The height of the top half at equidistant points, obtaining the sequence. After that, the bottom-half sequence was obtained.

The envelope was then divided into two halves again, and subsequently the left and right-hand sides were obtained through the geometric centre. As a result, two sequences were obtained. The feature vector sequence was composed of four dimensional vectors, the first component of these vectors was designed to help the HMM synchronization. A full explanation of these algorithms can be found in [7].

E. Hidden Markov Models (HMMs)

In this study, discrete HMMs [7] were selected to model each word's feature; this is to avoid making an assumption on the form of the underlying distribution. Each of the words (answers) was modelled with two left-to-right HMMs. The number of states in each signer's HMM words is thirty-five. This topology only allowed transitions between each state to itself and to its immediate right-hand neighbours. The classification, decoding, and training problems were solved with the Forward-Backward algorithm, the Viterbi algorithm, and the Baum-Welch algorithm.

The K-means algorithm was used for the training process to create multi-labelling VQ which made a soft decision about which code words were the closest to the input vector. To verify each answer, the log likelihood of the two HMMs that modelled the answer was obtained. The fusion of both scores can be performed by regarding the problem as a classification or a combination problem. If scores obtained were greater than the threshold, the answer was accepted.

The HMM software employed in this study was the GPDShmm toolbox which can be freely downloaded from http://www.gps.ulpgc.es/download/index.htm [8]. All the experiment performed were executed using Matlab in Windows 7 environment.

F. Experiment Evaluation Rates

The SAAS evaluations employed two rates, being classification and accuracy rates. The first rate, classification rate, was used to indicate the rate that the words in the testing datasets were recognised. The second rate, accuracy rate, was the rate which indicated the accuracy of the proposed system when the recognised words matched the answers to each of the questions.

III. Experimental Results and Discussion

This section reports results obtained from the experiments performed. As described earlier in Section II, there are three main types of dataset (CI, CII, and CIII) employed in the experiments.

Whereas CI contained only correct answers to the questions in its dataset, CII and CIII contained both correct and incorrect answers in their datasets.

The difference between CII and CIII was that CII did not use incorrect answers in the training process whereas CIII did. Each type of data was further divided into six sub-datasets; the results of employing each sub-dataset are described as follows:

a) Classification Rates Obtained from Employing CI (Trained with TR I and Tested with TE I).

The results of each sub-dataset are displayed in Table III. It can be seen from Table III that the best classification rate of 91.90% was obtained when the discrete HMMs were trained with 10% (300 samples) of the total dataset and tested with 90% (2,700 samples) of the total dataset.

TABLE III. Experimental Results of Each Sub-Dataset of the CI Dataset

Dataset: CI		
Sub-Dataset	**Training & Testing Ratio (%)**	**Classification Rat (%)**
S I	Train 80% - Test 20%	88.42
S II	Train 50% - Test 50%	89.50
S III	Train 40% - Test 60%	89.24
S IV	Train 30% - Test 70%	90.00
S V	Train 20% - Test 80%	90.34
S VI	Train 10% - Test 90%	**91.90**

It was also observed that the lowest classification rate of 88.42% was obtained when 90% (2,700 samples) of the dataset was employed for training and 10% (300 samples) of CI was used for testing.

High classification rates were expected as only correct answers to the questions were used in these experiments. From this dataset, it could be concluded that the classification rates tended to increase as the numbers of training samples were decreasing. This may result from a problem of overfitting from the larger number of samples used to train the classifiers.

b) Classification Rates Obtained from Employing CII (Trained with TR I and Tested with TE II).

In this dataset, the training dataset did not contain any of the incorrect answers to the questions, however, the testing dataset did. The results of each sub-dataset are displayed in Table IV.

TABLE IV. Experimental Results of Each Sub-Dataset of the CII Dataset

Dataset: CII		
Sub-dataset	**Training and Testing Ratio (%)**	**Classification Rate (%)**
S I	Train 80% - Test 20%	87.13
S II	Train 50% - Test 50%	86.31
S III	Train 40% - Test 60%	85.67
S IV	Train 30% - Test 70%	86.67
S V	Train 20% - Test 80%	87.10
S VI	Train 10% - Test 90%	**89.15**

It can be seen from Table IV that the best classification rate of 89.15% was attained when only 10% (300 samples) of the total dataset was used for training. This result was similar to the highest result of CI. Having incorrect answers to the questions in the testing dataset lowered the best classification rate by 2.75%.

It can be noted that the classification rates seemed to fluctuate more as the number of samples in the training datasets were decreasing; this was different from the results obtained when CI was used in the experiments (see Table III).

c) Classification Rates Obtained from Employing CIII (Trained with TR II and Tested with TE II).

In this dataset, the training dataset contained both correct and incorrect answers to the questions, the testing dataset also contained both correct and incorrect answers. The results of each sub-dataset are displayed in Table V.

TABLE V. Experimental Results of Each Sub-Dataset of the CIII Dataset

Dataset: C III		
Sub-dataset	**Training and Testing Ratio (%)**	**Classification Rate (%)**
S I	Train 80% - Test 20%	**89.88**
S II	Train 50% - Test 50%	88.24
S III	Train 40% - Test 60%	88.18
S IV	Train 30% - Test 70%	88.45
S V	Train 20% - Test 80%	**89.88**
S VI	Train 10% - Test 90%	89.12

From Table V, it can be observed that the best classification rate was increased when the discrete HMMs were trained with incorrect answers to the questions as well as the correct ones. The highest classification rate of 89.88% was obtained when the classifier was trained with either 80% or 20% (2,700 and 300 samples, respectively).

As expected, the classification rates increased when the incorrect answers were also used in training. The improvement went up to 0.73%. This 0.73% may appear to be nominal, however, in SAAS, this is very important. It may cause students to fail their exam if the system couldn't mark their paper correctly.

Similar to CII, small fluctuations across 3,000 samples could be seen as the numbers of samples in the training datasets decreased; this was different from the results obtained when CI was used in the experiments (see Table III). Since only small fluctuations were observed, this could not be considered statistically meaningful.

Consistent classification rates obtained from experiments performed on both CII and C III (see Table IV and V) suggested that the proposed SAAS system is robust with respect to the sizes of the training datasets.

It could be noted that by using CIII, the gap between CI (which only contained correct answers to the questions) was reduced to 2.02% compared to a 2.75% gap between CI and CII. The comparison results between each dataset's best classification rates, together with their corresponding settings, are displayed in Table VI.

TABLE VI. THE COMPARISON BETWEEN EACH DATASET'S BEST CLASSIFICATION RATE TOGETHER WITH ITS SETTING

Dataset	Training and Testing Ratio (%)	Best Classification Rate (%)
CI – S VI	Train 10% - Test 90%	**91.90**
CII – S VI	Train 10% - Test 90%	89.15
CIII – S V	Train 20% - Test 80%	89.88

It could be noted from Table VI, that the best classification rates were obtained when the numbers of training samples were small; from the experiments, the suitable range was when the HMMs were trained with 10 – 20% (300 – 640 samples) of the total dataset sizes of 3,000 and 3,400 samples, respectively.

DET curves of the experiments can be found in figure 3. Along the X-axis the FAR scores are plotted and along Y-axis the FRR scores.

d) Comparison between the proposed SAAS employing discrete HMMs and other off-line word recognition techniques found in the literature:

As discussed earlier under the Introduction Section, the amount of research conducted on off-line SAASs could be considered quite small; as a result the comparison in this study was performed with other off-line word recognition techniques found in the literature.

TABLE VII: THE COMPARISON BETWEEN DATASET SIZE (DS), CLASSIFICATION RATE (CR), AND ACCURACY RATE (AR) OF THE PROPOSED SAAS AND OTHER SYSTEMS FOUND IN THE LITERATURE

System – Feature Extraction Techniques	DS	CR (%)	AR (%)
English Numeral Recognition – Hybrid Features (Moment of Inertia and Projection) [8]	3,500	91.7	91.7
Arabic Handwriting Recognition System - baseline estimation – HMMs/MLP [9]	736	89.03	N/A
– Children's Handwritten Responses – HVBC FET [1] – Automated Assessment System - HVBC FET and constraints employed [2]	145 1,077	65.00 54.00	100 99.00
SAAS – G_GGF [10]	1,248	87.12	91.12
SAAS – G_WRL_MDGGF – SVMs [11]	3,400	94.88	98.09
The proposed SAAS – Geometrical Features - HMMs	**3,400**	**91.90**	**100**

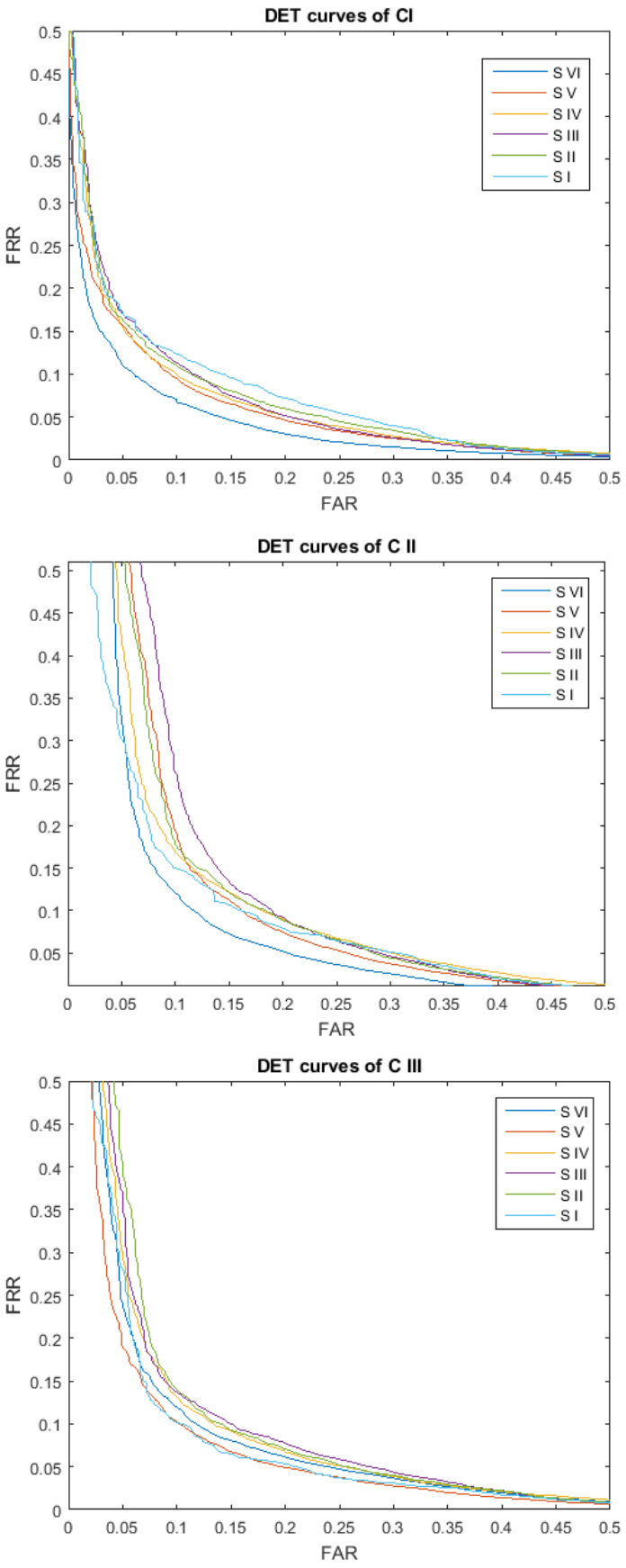


Fig. 3. DET curves of the experiments.

Upon observing the comparison in Table VII, it can be seen that when employing HMMs with the same dataset used in the previous study [11], the classification rate attained from this study is lower than [11] by less than 3%, however, it must be noted that since the proposed system is a SAAS, the accuracy rate is crucial. Any accuracy rate less than 100% would be considered unacceptable and unusable as students may fail their exam even though they answered the questions correctly. Given this reason, it can be considered that the proposed SAAS, employing HMMs, yielded better results. Furthermore, the proposed system only employed 10 – 20% of the total samples in the datasets for training compared to 80% used in the previous study [11] to achieve this result. Hence, it can be considered an efficient system.

The other existing SAASs were also able to achieve high accuracy rates of 99 – 100% [1] and [2]. However, the dataset sizes are relatively small (145 and 1,077, respectively). Furthermore, the classification rates were lower than the rate achieved in the present study (65%, 54%, and 91.90%, respectively). The proposed classification system can be considered comparable to some of those found in the literature [8], [9] and [10].

IV. Conclusion and Future Work

This study proposed a SAAS employing discrete HMMs; the geometrical features used were the envelope description and the interior stroke distribution in polar and Cartesian coordinates. Using HMMs, no explicit segmentation was required.

The experimental results were encouraging; the range of classification was from 85.67 – 91.90% with 100% accuracy. It must be noted that for SAASs, accuracy rates are crucial; any accuracy rates less than 100% would be considered unacceptable. Furthermore, by employing HMMs, the amount of samples required for training was reduced dramatically from 80% [11] to only 10 – 20%. This means that it is more efficient to employ HMMs in SAASs as the results are promising even though, only a small number of samples were required, and yet a 100% accuracy rate was attained.

Some suggestions are presented here for future work to improve the classification rates. Different algorithms and settings of HMMs can be applied. Furthermore, hybrid classifiers (e.g. HMMs & SVMs, HMMs & ANNs, etc.) can be employed. Rather than utilising a whole word recognition approach, segmentation-based recognition may be applied to SAAS. Different techniques such as deep learning can also be investigated on the datasets. More complex datasets (i.e. increasing from word to sentence level, larger dataset sizes, multilingual) can be collected and employed in future work.

Acknowledgement

The authors sincerely appreciate the volunteers who submitted their samples for datasets development, and those who helped in the collection process. Upon request, the database is available for download to the research community.

References

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

[2] Allan, J., T. Allen, N. Sherkat, and P. Halstead. “Automated assessment: it's assessment Jim but not as we know it”, In Proceeding of Sixth International Conference on Document Analysis and Recognition, p. 926-930, 2001.

[3] H. Suwanwiwat, U. Pal and M. Blumenstein, "An Automatic Off-Line Short Answer Assessment System Using Novel Hybrid Features," 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Gold Coast, QLD, 2016, pp. 1-8.

[4] G. R. Plamondon and S. N. Srihari. “Online and offline handwriting recognition: a comprehensive survey”, IEEE Trans. on PAMI, 22(1):63- 84, 2000.

[5] T. Plötz and G. A. Fink. “Markov models for offline handwriting recognition: A survey”, Int. Journal on Document Analysis and Recognition, 12 (4):269–298, 2009.

[6] M. A. Ferrer, J. B. Alonso and C. M. Travieso, "Offline geometric parameters for automatic signature verification using fixed-point arithmetic," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 993-997, June 2005.

[7] S. David, M.A. Ferrer, C.M. Travieso, J.B. Alonso, “gpdsHMM: A Hidden Markov Model Toolbox in the Matlab Environment,” CSIMTA, Complex Systems Intelligence and Modern Technological Applications, pp. 476-479, September 2004.

[8] 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.

[9] M. Rabi, M. Amrouch, and Z. Mahani “Contextual Arabic Handwriting Recognition System using Embedded Training based Hybrid HMM/MLP Models”, Transactions on Machine Learning and Artificial Intelligence, [S.l.], v. 5, n. 4, sep. 2017.

[10] H. Suwanwiwat, U. Pal and M. Blumenstein, "An Investigation of Novel Combined Features for a Handwritten Short Answer Assessment System," 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, 2016, pp. 102-107.

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