International Journal of Evaluation and Research in Education (IJERE)
Vol. 13, No. 4, August 2024, pp. 2388~2398
ISSN: 2252-8822, DOI: 10.11591/ijere.v13i4.28324

A paper-based cheat-resistant multiple-choice question system with automated grading

Lienou T. Jean-Pierre[1], Djimeli-Tsajio Alain Bernard[2], Noulamo Thierry[3], Fotsing Talla Bernard[3]
[1]Department of Computer Engineering, College of Technology, The University of Bamenda, Bamenda, Cameroon
[2]Department of Telecommunication and Network Engineering, University of Dschang, Bandjoun, Cameroon
[3]Department of Computer Engineering, University of Dschang, Bandjoun, Cameroon

Article Info

Article history:

Received Nov 30, 2023
Revised Jan 18, 2024
Accepted Jan 31, 2024

Keywords:

Cheat resistant assessment
Dissimilarity measure
K-mode clustering
MCQ
Reinsertion distance

ABSTRACT

This paper focuses on how to reduce cheating and minimize errors while automatically grading paper-based multiple-choice questions (MCQ) by making the whole process relatively fast, less expensive, more credible, and fairer especially when the number of examinees and number of questions are large. Credibility is obtained when techniques and best practices are introduced in the design process of MCQ. Fairness is obtained by personalizing evaluation through permutation of answers and questions. The distance introduced in personalization has led to the modification of the traditional automatic grading process where an application mapping the test number with its responses in the grading software is loaded automatically at each start of the grading process. On the extracted header fields, 2DFFT is applied as well as the reduction of computed coefficients to obtain the corresponding final local characteristic in the representation. The minimization of image processing errors is then obtained by training a support vector machine (SVM) for handwriting optical character recognition (OCR) using the Mixed National Institute of Standards and Technology (MNIST) dataset with 99.5% accuracy. The tests are carried out in several subjects at Fotso Victor University Institute of Technology (UIT) in Bandjoun and the ColTech of the University of Bamenda and teachers as well as students after investigation have confirmed that our method reduces cheating and improves the error rate during grading with fewer complaints.

Corresponding Author:

Lienou T. Jean-Pierre
Department of Computer Engineering, College of Technology, The University of Bamenda
Bamenda, Cameroon
Email: lienou@gmail.com

1. INTRODUCTION

The design of credible and fair multiple-choice questions (MCQ) associated with automated grading is not evident, especially when it comes to paper-based type. For an online assessment, there are fewer problems when it comes to automating the marking process. However, with a very large number of students, in certain environments where electricity is unstable, quality of poor communication bandwidth, and the network infrastructures (servers, software, connection capacity of access points) do not follow the latest technologies, it is recommended to use paper-based MCQ. The grading can be done using a smartphone that can help transfer the marks to a database on a local server. The growing number of students forces assessments to be made with students sitting tightly together and this increases the possibility of cheating.

Several research works have been carried out on the different collusion techniques that students use to cheat during MCQ-type assessments on paper or online-based using image processing techniques and in particular convolutional neural networks (CNN) that are still in full research [1], [2]. The grading equipment used at the University of Bamenda and the University of Dschang is of the "fiber laser marking machine" type which requires precise paper weight, specific coloring ink, the uniqueness of the test and the answer. Another factor is education malpractices where examinee can be looking at copies of the neighbors; this can be curbed by using the permutations of the questions and the answers of the test. This situation is already properly handled online and to the best of our knowledge, it is not yet feasible for paper-based assessments that are graded automatically.

An ideal system must have the test with questions also swapped as online MCQ and also must be easily markable. To address this issue, as per paper-based, a software layer designed for this purpose is inserted between the capture and grading steps in the automatic marking tool. Literature in the domain does not provide benchmarks for the personalization of tests and we opt for a permutation of questions and a calculation of dissimilarity between tests which will make it difficult for students to compare questions while sitting side by side. The distance between questions of the same number varies in a given set of students called here cluster set. Not only is the filtered permutation a difficult exercise, but the approach takes into account the layout of the students in the seating space, but the concept of distance introduced is close in literature to that of Manhattan, Hamming, and reinsertion distance at the same time [3]. If such system is not built, many paper-based assessments will end up being not fair, not credible and will not fully benefit from the advances noticed in the information technology field.

In the exam database, each candidate's name is associated with a special code called an anonymity number which is pending the corresponding grade. This anonymity number is associated with the candidate's copy-labeled registration number. During automatic correction, this registration number field is extracted and the corresponding handwriting digits are digitized to match the candidate's name and grade in the information system. Although handwriting digit recognition is an old problem [4], automatic and accurate correspondence between the candidate mark and name is a challenge as an anonymity number permutation would result in a permutation of the corresponding candidate's grade and a duplicate of the anonymity number would lead to a verification process that could be long and tedious, with manual correction. For a classifier to be able to generate a good discriminative model, the representation of the input space is of key importance.

Many classification algorithms cannot achieve the expected performance when faced with difficult real-world problems, because most of these groups of classifiers are inherently incapable of transforming their input space to gain class separability. A popular approach to classification today is to use a pre-trained CNN, to extract useful and informative features of images and use it as a starting point for training, assuming that the source and target domains are related to each other [5]–[7]. Su *et al.* [8] has proposed a face recognition method that combines both global and local discriminative features. Global features are real and imaginary components of a low frequency of the image 2D Fast Fourier Transform (2DFFT). The Fourier transform helps to recognize position-shifted characters in the magnitude spectrum. Fourier transforms and their variants such as Fourier moments, Fourier descriptors, local binary pattern Fourier histogram, polar Fourier transform, Fourier-Mellin Transform, or wavelet-Fourier descriptors are widely used for feature extraction and shape classification [9], [10]. Many works have also been done in pattern recognition and there are still challenges. A review of such works is done in [11]–[13].

The design of MCQs must follow rules to have a credible evaluation. Several works have been carried out for the credibility and fairness of knowledge checks for both online and paper-based assessments. The result is the "Best Practices" proposal [1], [12]. For example, assessing mastery of the concept of multiplication for an online assessment with macros can have fairness issues and on paper, automatic correction becomes a complicated problem when the answers are shuffled [14]. The corresponding code in an algorithmic language is shown in Algorithm 1.

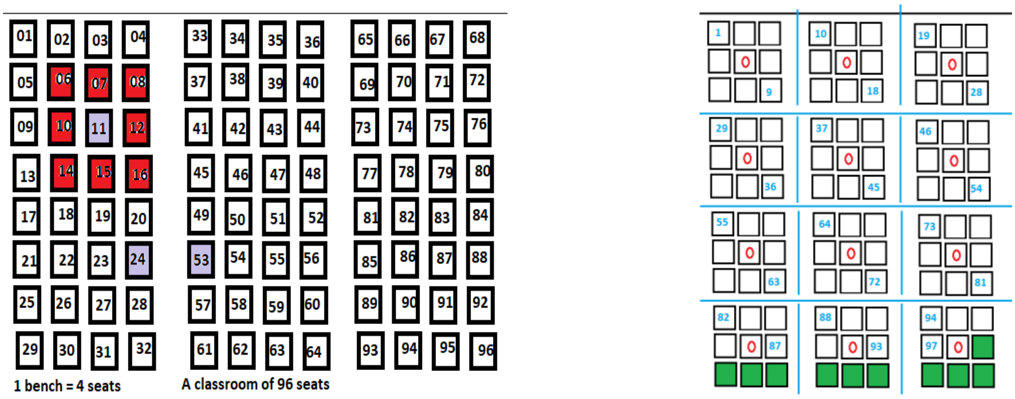
Algorithm 1. Macros for personalization of a question on multiplication
```
$a=randbetween(10, 99);
$b= randbetween(10, 99);
$c=$a*$b;
```

Then, 5 answers are proposed including $a*$b. This situation can generate for one student 10*10=100 and for another 87*93=8091 which are of different complexity and would not only introduce inequity into the grading process but could make automatic marking nearly impossible. When the choice of questions is made from a question bank with best practices criteria that must be respected, such as that of questions of complexity of an equivalent level, this can become an extremely complex problem. A study by Nguyen *et al.* [15] have proposed a multi-swarm optimization for the extraction of questions in a bank to form tests of equivalent complexity by proposing a parallel version of the algorithm to reduce calculation times. Costello *et al.* [16] have made an in-depth study of the situations that can lead to cheating in MCQ-type assessments in MOOCs. It emerges that solutions of the "all of the above" type do not allow good answers permutation and make the evaluation not very credible. McKenna [17] studied the different techniques that students can use to find the right answers without knowing the right answers with high probability. One of the solutions that we find is the personalization of the tests.

Taking random questions from a question bank to constitute the test paper can make the evaluation unfair. Our suggestion is to use enough questions in the test and swap the same questions to form the other tests. For the "online" versions, a lot of work has been done and implemented in various online training platforms and there is no difficulty for automatic marking. However, on the paper-based versions, the grading process steps need to be adjusted. Bankar *et al.* [1] proposed an algorithm to permute answers in MCQs. The architecture of our system certainly reduces the necessary personnel and has the same actors as that of a traditional remote platform, namely an administrator, a teacher, a student, and a server. For a fair assessment (the same questions for all students), customization by swapping the position of the questions can be a credible solution. The distance between the questions of two students sitting side by side therefore becomes an important parameter. The proposed technique is close to the one of Cicirello [3].

The "distance reinsertion" applicable in clusters for deep learning and derivative works allows for several types of permutation [18], [19]. At the end of the implementation of some algorithms, recursive or not, sometimes based on exchanges and distances are not respected so that two neighbors cannot easily cheat. Research by Shaikh *et al.* [12] have proposed the best architecture at the grading level, one based on CNN to automate the grading process. The framework proposed by Balaha and Saafan [20] also solves technical and administrative challenges that occur. A study presented an algorithmic framework for exhaustively generating generic rectangulations, and diagonal rectangulations in constant time in the worst case [21]. Several researchers studied dissimilarity in a k-mode cluster algorithm [22]–[24]. Other comparative studies have been carried out for the dissimilarities between clusters of categorical data [25]–[27] or for the analysis of simple matching dissimilarity measure (SMDM), distance learning in categorical attributes (DILCA), domain value dissimilarity (DVD) and simple weighted matching dissimilarity measure (SWMDM) criteria [23]. Zhou *et al.* [24] aggregate a set of k-node dissimilarity measures of clusters.

The problem of fair evaluation is not a problem of only higher education but also that of elementary schools in Cameroon [28]. An MCQ is a test having a set of questions. Each question has a certain number of answers. A subset of answers forms the correct answer that corresponds to a certain mark. A hall is a venue where a test takes place. A hall has a series of benches in columns and rows. There may be some spacing to let invigilators pass or to let students pass and go to their seats that can be numbered. A set of rows of one column forms a cluster or module. Two columns can have two different numbers of rows. For simplicity, each seat is numbered and the clusters have their cartesian coordinates in terms of row and column. This concept is explained in Figure 1 with an example of seat labeling. Figure 1 (a) depicts the seat labeling column first and the eight connected neighbors, and Figure 1 (b) shows the cluster formation with the labeling row first. We also provide a simplified approach that improves the performance of classical handwriting digits classifiers by providing an overview of its behavior with an entry presented as a local structure of the target entry in the frequency domain. Applications are made on the support vector machine (SVM) for the extraction of anonymity numbers during the process of automatic correction of MCQ. As a reminder, we recap some differences and similarities between online assessment and paper-based challenges in Table 1.



(a) (b)

Figure 1. Seats labelling for (a) column first and the eight connected neighbors and (b) clusters formation with labeling row first

Table 1. Advantages and disadvantages of online and paper-based MCQ with automatic grading

	Online with automatic grading		Offline (Paper-based with automatic grading)	
	Advantages	Disadvantages	Advantages	Disadvantages
Question pooling	Easy to build a quiz	Considerable efforts are needed to guarantee fairness	Contributes to reducing answer copy	Quizzes are printed individually Fairness not guaranteed
Randomize question display	Easy to configure	Randomize algorithms may not be efficient	Reduces cheating if enough questions	Careful design is needed
Shuffle the answer options inside a question	Easy to configure	Some answers should be fixed	Contributes to reducing answer copy	None
Add a time limit	Easy to configure	None	Contributes to reducing answer copy	None
Proctoring	Screen and microphone-sharing tools are expensive	None	Invigilators are easy to deploy; Reduce many forms of cheating	Invigilators need to be paid
Disable tab switching	It wastes time for those trying to look for answers somewhere	Very limited advantage since other browsers may be used to look up Answers	Not Applicable	Not Applicable
Privacy & security controls	Intellectual property protected	None	Unless paper questions are repeated	The invigilator needs to be careful and identify the examinee
Require ID	Many LMSs do not implement	None	Maps anonymity number	The OCR process is not 100 % error-free

Paper-based MCQ is considered the most appropriate for many circumstances even though it needs a lot of investment at certain point of time. This paper thus proposes an MCQ evaluation approach that minimizes image processing errors while making evaluation more credible and fairer with error mapping reduction by using deep learning technics in computer vision for the grading process. The rest of the paper exposes the experimental environment, our permutation algorithm, and system architecture.

2. RESEACRH METHOD

2.1. Architecture of the new system

The system is sketched in Figure 2. The boxes of the figure have the following roles. In step A, the evaluator selects the questions to form the test from the question bank. Depending on the organization and the pedagogical goal, a test covers a subset of concepts to be evaluated. In step B, the set of questions is known. They are numbered and the correct answers are identified for each question. Let's suppose a question has five answers (A, B, C, D, E), each answer having a positive or negative mark if it is checked with two correct ones (B, D). In the system, we save ABCDEBD. Everything that is repeated is supposed to be ticked to have the full answer. The first 4 columns of Table 2 show an example. In step C, the correct answer to the typical test is saved in the database. In step D, the hall seats are numbered before the exam. We use a cartesian system. Figure 1 (a) shows two levels of axis. The macro level in terms of columns here is 3 and in terms of rows 3.

The macro levels axis is denoted by the capital letter RC. The micro level is denoted by rows and columns also rc. The modularity chosen is done in such a way that examinees from a group cannot have the same order of questions. Questions are distant from each other so examinees will think that the questions are personalized. Meanwhile, all the students were examined with the same questions. Seat number 63 has the following coordinates RrCc=R3r2C2c2. The macro levels can go up to any level meanwhile the micro level defines the minimum and the maximum number of examinees in the cluster. So, for C here, we have $1 \leq R \leq 3$. This is the same for the columns. The micro level is formed by the number of layers we have fixed ourselves to accept that there will not be a possibility to glance and copy the results. As for Figure 1 (a), we used the 8 connected points that will always match the formula $(2n+1)^2$ the number of seats in a layer. For Figure 1 (b), the number of seats is fixed to 3 rows. So, per row, we have $1 \leq c \leq 4$ and $1 \leq r \leq 3$. This gives the formula $r*c$ is the total number of seats per cluster. Without loss of generality, the module gives an indicator of which questions will be distant within the cluster. For clarity of our methodology, we use negative and positive directions by introducing the concept of center. The center in a module has the coordinates (r=0 and c=0). Their macro coordinates will then vary. If a module has X seats on an axis, any x should have its coordinate along that axis verifying the relationship.

$$\left[-\,{}^{(X-1)}\!/_{2} \leq x \leq +\,{}^{(X-1)}\!/_{2}\right] \tag{1}$$

For instance, if X=6, any x will have its coordinate along that axis between [-3, 2], and the center is 0. It goes the same for the row and column. If X=7 seats, $x \in [-3, +3]$ with the center at 0. This is important to compute the distance concept we are to introduce. For the case of the radius n=2, the number of seats in a cluster is 25. We suppose that questions must have a distance of 25.

$$D_{min} \geq (2n+1)^2 \quad (2)$$

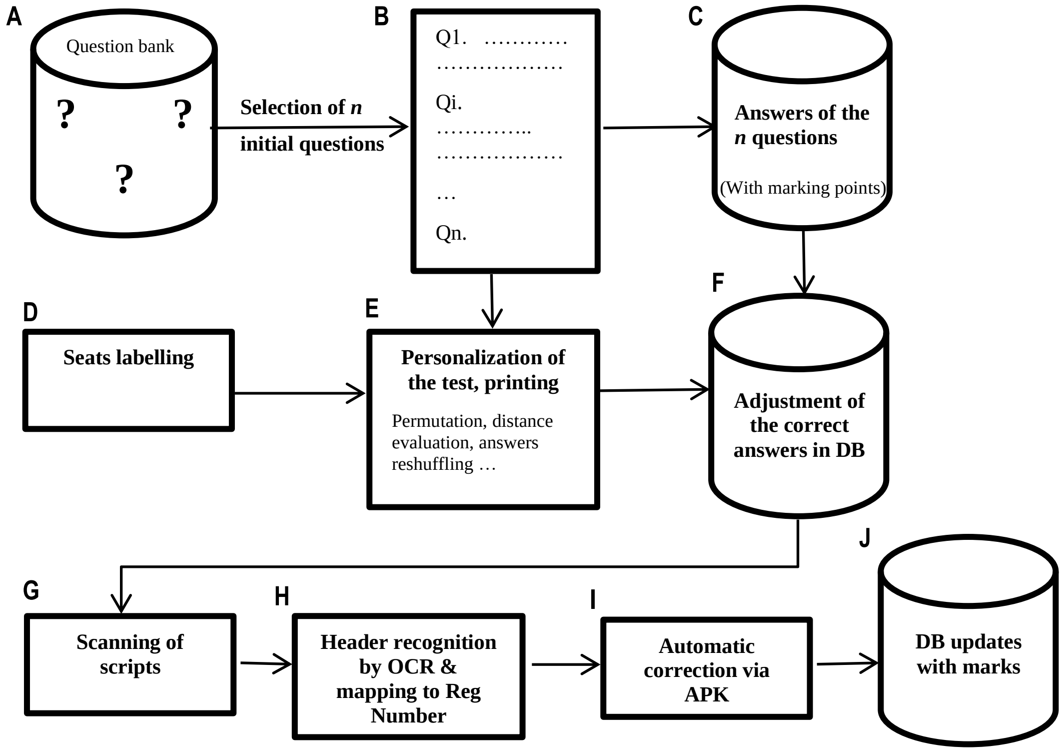


Figure 2. Steps in paper-based automatic MCQ evaluation

Table 2. Steps of mapping one question with reshuffled answers

Question number	Marks	Answer	% answers	New question	Marks	Old answers	New answers reshuffled	Answers with corrections in IS	% answers
Question 1	3	A	0%	Question 29	3	A	D	A	60%
		B	40%			B	B	B	40%
		C	-25%			C	A	C	-25%
		D	60%			D	C	D	0%
		E	0%			E	E	E	0%

In step E, all the questions of the test are then saved in a new table of the database. At this point, question 1 of the new test is question 29 for the examinee sitting nearby. The paper question personalization goes here with all the possible acceptable (distance greater than a threshold) permutations. In step F, a table in the database of all the mapping solutions of the regenerated and reshuffled questions is created: Test number, order of questions, and adjusted answers. In step G, the test is administered, the scripts are collected and it is time to scan or take snapshots from a smartphone. In step H, one has the image scanning of the papers collected from the hall. An OCR process maps the question number with the student registration number. In step J, the grading process is done and the grade is transferred to the database.

Table 3 is done for nine cases using Algorithm 2 for general permutation and distance calculation [21], [26], [29], [30]. The steps from A to F are pre-examination tasks and from G to I are post-examination ones. In the pre-examination tasks, steps B and C have the model of the designed database, step E corresponds to a new algorithm and a new distance calculation between questions, and step I generates the *.apk file which corresponds to the software to be loaded in the mobile device be it a smartphone or a tablet.

In Algorithm 2, we present an excerpt of the algorithm of the automatic permutation. All the generated permutations are stored in a database and later extracted for dissimilarity evaluation. A subsequent permutation is discarded if the dissimilarity has not reached the fixed threshold. It is computed as we form a group of arrays in linear codes except that the distance used here is the reinsertion distance algorithm. Table 3 shows the presentation of manual permutation satisfying the dissimilarity criteria that is fixed based on the number of cluster sets and the number of seats, a student cannot see and read beyond that distance.

Table 3. Permutation results before swapping

No.	RV1	RV2	RV3	RV4	RV5	RV6	RV7	RV8	RV9
1	42	8	18	66	53	2	20	56	72
2	27	51	6	56	11	8	46	65	11
3	21	6	2	44	65	10	10	33	12
4	23	4	66	70	18	69	41	27	27
5	55	27	10	4	1	49	65	14	31
6	48	28	29	28	36	12	35	64	53
7	68	25	1	41	25	51	70	37	48
8	45	52	4	37	47	40	29	7	54
9	72	5	41	13	56	36	34	60	39
10	49	36	5	61	5	7	74	20	66
11	9	26	69	46	21	6	56	53	33
12	32	66	27	19	58	58	50	31	59
13	24	61	63	23	69	57	32	34	37
14	22	75	25	50	38	17	72	61	45
15	73	10	7	27	57	52	73	71	44
16	11	15	38	60	70	60	59	38	56
17	2	34	55	48	62	54	39	62	21
18	30	62	34	20	39	71	66	6	18
19	26	74	56	57	8	73	1	45	2
20	1	73	68	7	15	34	4	67	70

Algorithm 2. Recursive version of permutation
```
void heapPermutation(int a[], int size, int n)
{
    if (size == 1) {
        printArr(a, n);
        saveArr(a,n);
        return;
    }
    for (int i = 0; i < size; i++) {
        heapPermutation(a, size - 1, n);
        if (size % 2 == 1)
            swap(a[0], a[size - 1]);
        else
            swap(a[i], a[size - 1]);
    }
}
```

2.2. Extraction of anonymity number field

Feature representation in the frequency domain is a preprocessing approach that transfers local spatial information into its correspondent in the frequency domain. Because we are concerned with the local structure of the image a window size (F) is needed. For compliance, we will only use square windows to preserve the relationship between the pixels by learning the image features using said small squares of input data. For the representation, the window is then shifted until it covers the entire image while computing the 2D Fast Fourier Transform (2DFFT). If we start with the window in the upper left corner of the input table and then drag its global locations down and to the right, we can move our window more than one pixel at a time ignoring the intermediary locations. We refer to the number of rows and columns traversed per slide as the stride (S). The stride is a parameter that tells how many pixels the window is shifted. For example, if we use the 6x6 image with a stride of 2 and a window of size 3, we end up with the Local Fourier Coefficients presented in Figure 3. Figure 3 (a) shows the first 2x2 coefficient extracted and Figure 3 (b) shows the LCF extracted after the second shift.

In this illustration, the square window of size 3 is superimposed top left of the image so that the maximum number of window boxes is used before the computation of the 2DFFT. Because we are interested in the preservation of the image energy, only the magnitude of the 2DFFT is reported. The window is then shifted successively by two pixels to the right and successively by two pixels (stride) downwards to obtain the required matrix that will undergo normalization before being used. The number of Fourier coefficients for the illustration given in Figure 3 (a) is (10x10). By remembering that the complexity of the calculation of the 2DFFT coefficients increases with the size of the window, one can note the explosion of the number of Fourier coefficients when the stride tends towards 1. The Fourier transform is a window-specific task where one would like to keep higher frequencies which represent the edges of the image and lower frequencies which represent the details of the image while ignoring frequencies that correspond to the homogeneous area of the image. These frequencies, which give rise to non-crucial information, are located in the middle of the widths and lengths of the representation of the coefficients in the window.

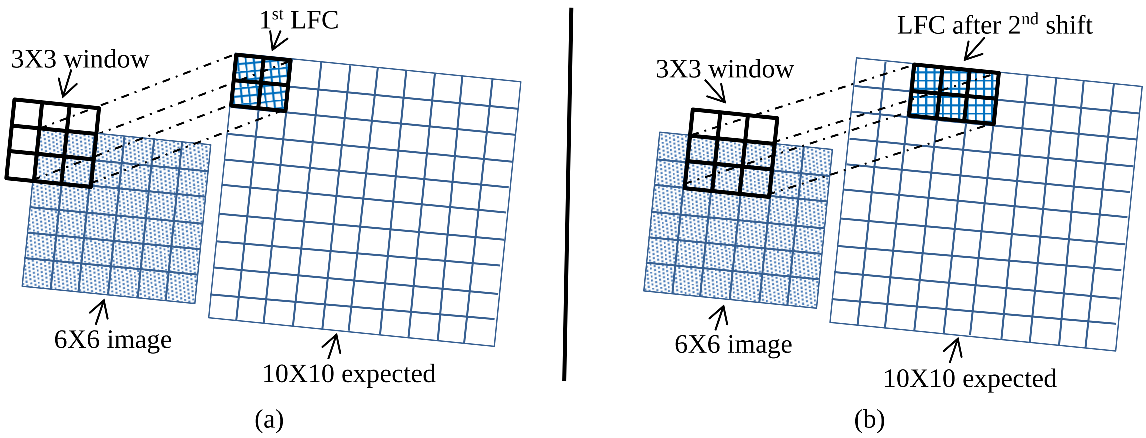


Figure 3. Principle of Local Fourier Coefficients (LFC) extraction on Image for 3x3 window and stride=2 of (a) the first 2x2 coefficient extracted and (b) the LCF extracted after the second shift

Support vector machine (SVM) classifier was used in this work. SVM for classification uses hyperplanes for decision boundaries in the input space or in the high-dimensional feature space from a labelled training dataset. Throughout the training phase, SVM takes each element in a labelled data matrix and treats it as a row in an input space or a high-dimensional feature space, where the number of attributes identifies the dimensionality of space. Multi-class SVM includes several two-class subproblems that can be easily combined using one-over-all and one-over-one coding algorithm [31], [32]. In this case, we applied the Matlab *fitcecoc()* function on the LFC features for the classification task.

The accuracy increases with the size of the window and stabilizes for window value 15 before decreasing to a small value. To explore smaller strides for higher window size where the number of computed LFCs is very high, we explored the behavior of computed error concerning the percentage of LFC removed and saw that there is no significant effect on classification error when the coefficients are reduced. The maximum accuracy of 99.51 obtained is state-of-the-art. Although works using CNNs are becoming state-of-the-art techniques for the classification of handwriting digits with interesting accuracy [4], CNNs are slower than classical classifiers, the training process is CPU-intensive and time-consuming. Some characters were not recognized by the system. To assess the degree of satisfaction with our recognition system concerning unrecognized characters, the labels of the latter were masked and the handwriting was submitted to a group of 50 teachers for labelling.

3. RESULTS AND DISCUSSION

3.1. Results

For the questions number of nearby sitting examinees, a minimal distance concept is introduced and it can be measured by using (2). Two permutations techniques are proposed when the number of questions is enough to keep the questions distant each other from test number to another test number. The manual technique is shown in Table 3 and the automatic one in the permutation code in Algorithm 2. A program is being developed to compute the permutations and filter the acceptable ones. The concept of cluster introduced using (1) permits to compute how far the test number could be repeated. The reshuffle of the answers on specific question having X answers produces about X different set of answers for the same question. The standard used was X=4 which produces 24 different permutations of the answers. Considering the distance between questions, our standard was to use 8-connected students to determine the minimal distance between two questions. The experiment was made with n=1 as shown in (2). Therefore, the minimal distance between questions is 9.

It was asked to assign a label to the image of the characters presented in Figure 4 knowing that they are numbers ranging from 0 to 9. After counting, we realized that the results were very variable from one volunteer teacher to another and that everyone had their own understanding of said characters. We obtained a good labelling rate ranging from 17/24 to 6/24 with an average of 11.6/24. These rates sufficiently show the polysemic character of the handwriting not recognized by our system. One of the merits of our system is not to recognize the characters whose recognition and interpretation remain ambiguous for the human being since machine learning is only the transfer of human knowledge to the machine. Data augmentation techniques are generally used to increase the amount of data by combining it with slightly modified copies of previously existing data. Then, augmented data can include biases such as those present in the testing dataset. As a result among others, it resolves class imbalance issues in classification and improves the accuracy of model prediction. Therefore, comparing the results obtained with the data augmentation techniques with the results of those who do not use it, is not fair and we have chosen not to compare the results of the proposed approach with those of the literature using data augmentation techniques for MNIST handwriting digits dataset. Table 4 presents the most competitive results (error rate<1%) found in the state of the art including the proposed approach for the MNIST dataset without data augmentation and CNN [4].

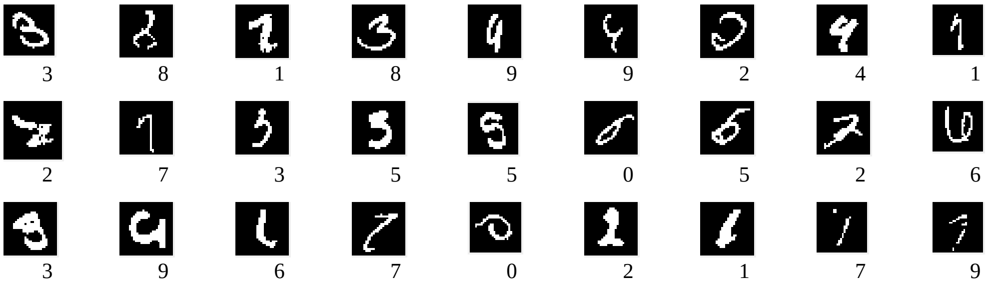


Figure 4. Half of the character’s images are not recognized by our system and their label

Table 4. Comparison of the results of our studies with those found in the literature for the MNIST database without data augmentation and CNN [4]

Technique	Test error rate	Technique	Test error rate
NN 6-layer 5,700 hidden units	0.35%	Pooling + SVM	0.64%
MSRV C-SVDDNet	0.35%	Virtual SVM, deg-9 poly, 1-pixel jit	0.68%
Committee of 25 NN 2-layer 800 hidden units	0.39%	NN 2-layer 800 hidden units, XE loss	0.70%
HOPE+DNN with unsupervised learning features	0.40%	SOAE-σ with sparse connectivity and activity	0.75%
Proposed approach (Local Fourier Features and SVM)	0.49%	Deep convex net	0.83%
K-NN (P2DHMDM)	0.52%	CDBN	0.82%
COSFIRE	0.52%	S-SC + linear SVM	0.84%
K-NN (IDM	0.54%	2-layer MP-DBM	0.88%
Virtual SVM, deg-9 poly, 2-pixel jit	0.56%	DNet-kNN	0.94%
RF-C-ELM, 15,000 hidden units	0.57%	2-layer Boltzmann machine	0.95%
PCANet (LDANet-2)	0.62%	NN 2-layer 800 hidden units, MSE loss	0.90%
PCANet (LDANet-2)	0.62%	DNet-kNN	0.94%
K-NN (shape context)	0.63%	2-layer Boltzmann machine	0.95%

Figure 5 shows an example of an answer sheet used in the College of Technology of The University of Bamenda in Cameroon where the answer sheet is printed at the same time as the question paper is being printed to track the paper number. This number is important to load the correct answers corresponding to that question paper number. The only registration number fields filled manually permit to mapping of the course code and the examinee's registration number. The permutation process of questions to personalize the paper question can be done manually if the total number of asked questions is small. The result of computation done on a powerful machine can be reinserted in the software to avoid long permutations that can crash down small systems since the number of permutations in a recursive procedure may bring a stack overflow. Once the developed apk is installed on an Android phone, the snapping of answer paper sheets can start. The user selects the mode that can be a batch mode (all the answer sheets are snapped at a time and put in a folder) or an interactive mode where the marking process starts by loading the corresponding paper number question. The Registration number is then mapped to the database with the obtained mark. An Excel file can also be generated when it is in batch mode. The proposed MCQ evaluation system that minimizes image processing errors while making evaluation more credible and fairer is used in Fotso Victor University Institute of Technology and ColTech for several subjects. A statistical study from teachers as well as from students shows that our method reduces cheating and improves the error rate during grading with fewer complaints.

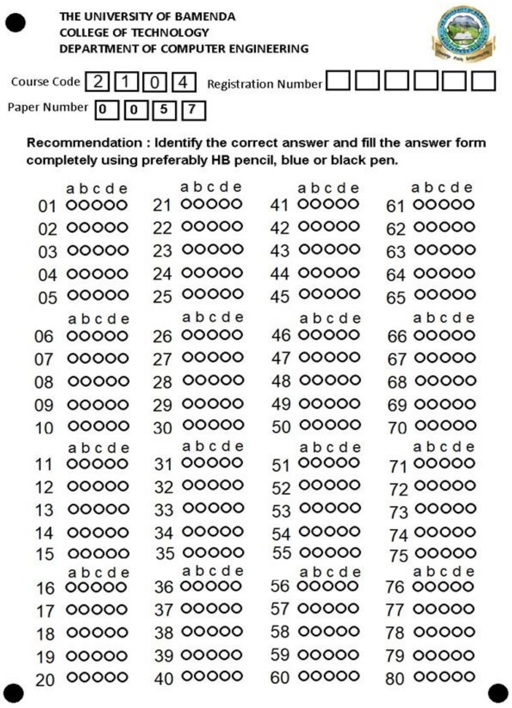


Figure 5. Sample of Answer sheet in ColTech

3.2. Discussion

Many machine learning models exist and were not all tested in our case to select the best one. No clear heuristic to the best of our knowledge can help fine tune the best selected model in terms of the best activation function, and the loss function to be used [31], [32]. At the permutation level, it would be better to look for a function that could map the i^{th} question to a j^{th} question depending of the sitting position of the examinee. This case will speed up the process of grading since the mapping indicates how to get the right answers instead of reading the corrections file from the disk for each question paper. Furthermore, it will reduce the computation complexity needed to do the acceptable permutations if the number of questions is large.

4. CONCLUSION

This paper proposes a simple method to reduce cheating during MCQ paper-based evaluations in a crowded hall without reducing fairness and also ameliorating the grading process. The method used is to personalize the question paper by reshuffling the order in which exam questions appear in the paper. We took into account the geographical position of the examinee’s seating place and also reshuffled the carefully built answers with credible distractors. A formula is provided to let the exam administrators know the granularity or the size of a cluster of students to generate the number of different question papers and what is the minimal distance for a question to another for examinees in the same cluster or neighboring clusters. It is therefore possible to figure out where question paper numbers can be repeated without significantly affecting the possibility of doing common work amongst examinees. The features of the MNIST handwritten digit dataset extracted with 2DFFT are used to train the SVM and the percentage of error in the recognition process is about 0.49% which outperforms many results in the literature. The whole system has two parts: One is installed on one server that is responsible for calculating the permutations satisfying the distance criteria to generate the paper questions. The retained permutations are stored and at the same time the database of students and their grades. The other part is installed on a tablet or Android phone and is responsible for converting the answer sheet into a jpeg image. The software then recognizes the owner's answer sheet, grades it, and generates an Excel file that can be easily imported into the database.

The implications of the contributions are many. First of all, there is no need to maintain or buy new optical-mechanical and electronic grading machines that use specific answer papers to be filled with specific ink. The educational system is more credible if implemented and the quality of education is improved. The perspectives of the work may be either to look for a possibility to design a function that can find a bijection to map the question paper to the corresponding answers so we do not have to load answers for all the questions or, to design a low-cost equipment on which we put all the answer sheets to snap at a high speed.

REFERENCES

[1] M. Bankar, P. Bhor, P. Bhalerao, and P. A. Dere, "Automated generation of question paper for online MCQ test by using Shuffling algorithm," *International Journal for Research Trends and Innovation*, vol. 3, p. 219, 2018.

[2] S. Manoharan, "Cheat-resistant multiple-choice examinations using personalization," *Computers & Education*, vol. 130, pp. 139–151, Mar. 2019, doi: 10.1016/j.compedu.2018.11.007.

[3] V. A. Cicirello, "Classification of permutation distance metrics for fitness landscape analysis," *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, vol. 289, pp. 81–97, 2019, doi: 10.1007/978-3-030-24202-2_7.

[4] A. Baldominos, Y. Saez, and P. Isasi, "A Survey of Handwritten Character Recognition with MNIST and EMNIST," *Applied Sciences*, vol. 9, no. 15, p. 3169, Aug. 2019, doi: 10.3390/app9153169.

[5] A. M. Hassan, M. B. El-Mashade, and A. Aboshosha, "Deep learning for cancer tumor classification using transfer learning and feature concatenation," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 6, pp. 6736–6743, 2022, doi: 10.11591/ijece.v12i6.pp6736-6743.

[6] M. Ghafoorian *et al.*, "Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation," in *MICCAI 2017: Medical Image Computing and Computer Assisted Intervention − MICCAI 2017*, 2017, pp. 516–524, doi: 10.1007/978-3-319-66179-7_59.

[7] M. Huh, P. Agrawal, and A. A. Efros, "What makes ImageNet good for transfer learning?" Computer Vision and Pattern Recognition, 2016, [Online]. Available: http://arxiv.org/abs/1608.08614.

[8] Y. Su, S. Shan, X. Chen, and W. Gao, "Hierarchical Ensemble of Global and Local Classifiers for Face Recognition," *IEEE Transactions on Image Processing*, vol. 18, no. 8, pp. 1885–1896, Aug. 2009, doi: 10.1109/TIP.2009.2021737.

[9] K. K. Thyagharajan and I. Kiruba Raji, "A Review of Visual Descriptors and Classification Techniques Used in Leaf Species Identification," *Archives of Computational Methods in Engineering*, vol. 26, no. 4, pp. 933–960, Sep. 2019, doi: 10.1007/s11831-018-9266-3.

[10] K. Hameed, D. Chai, and A. Rassau, "A comprehensive review of fruit and vegetable classification techniques," *Image and Vision Computing*, vol. 80, pp. 24–44, Dec. 2018, doi: 10.1016/j.imavis.2018.09.016.

[11] O. Buza, "Automatic Tests Correction System in Education," *2022 23rd IEEE International Conference on Automation, Quality and Testing, Robotics - THETA, AQTR 2022 - Proceedings*, 2022, doi: 10.1109/AQTR55203.2022.9801988.

[12] E. Shaikh, I. Mohiuddin, A. Manzoor, G. Latif, and N. Mohammad, "Automated grading for handwritten answer sheets using convolutional neural networks," *2019 2nd International Conference on New Trends in Computing Sciences, ICTCS 2019 - Proceedings*, 2019, doi: 10.1109/ICTCS.2019.8923092.

[13] A. M. Tavana, M. Abbasi, and A. Yousefi, "Optimizing the correction of MCQ test answer sheets using digital image processing," *2016 8th International Conference on Information and Knowledge Technology, IKT 2016*, 2016, pp. 139–143, doi: 10.1109/IKT.2016.7777754.

[14] H. O. Bendulo, E. D. Tibus, R. A. Bande, V. Q. Oyzon, M. L. Macalinao, and N. E. Milla, "Format of Options in a Multiple Choice Test Vis-a-Vis Test Performance," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 6, no. 2, p. 157, 2017, doi: 10.11591/ijere.v6i2.7594.

[15] T. Nguyen *et al.*, "Multi-Swarm Optimization for Extracting Multiple-Choice Tests from Question Banks," *IEEE Access*, vol. 9, pp. 32131–32148, 2021, doi: 10.1109/ACCESS.2021.3057515.

[16] E. Costello, J. C. Holland, and C. Kirwan, "Evaluation of MCQs from MOOCs for common item writing flaws," *BMC Research Notes*, vol. 11, no. 1, p. 849, Dec. 2018, doi: 10.1186/s13104-018-3959-4.

[17] P. McKenna, "Multiple choice questions: answering correctly and knowing the answer," *Interactive Technology and Smart Education*, vol. 16, no. 1, pp. 59–73, Mar. 2019, doi: 10.1108/ITSE-09-2018-0071.

[18] H. Ibrahim, Z. Omar, and A. M. Rohni, "New Algorithm for Listing All Permutations," *Modern Applied Science*, vol. 4, no. 2, Jan. 2010, doi: 10.5539/mas.v4n2p89.

[19] R. Sedgewick, "Permutation Generation Methods," *ACM Computing Surveys (CSUR)*, vol. 9, no. 2, pp. 137–164, 1977, doi: 10.1145/356689.356692.

[20] H. M. Balaha and M. M. Saafan, "Automatic Exam Correction Framework (AECF) for the MCQs, Essays, and Equations Matching," *IEEE Access*, vol. 9, pp. 32368–32389, 2021, doi: 10.1109/ACCESS.2021.3060940.

[21] A. Merino and T. Mütze, "Combinatorial Generation via Permutation Languages. III. Rectangulations," *Discrete & Computational Geometry*, vol. 70, no. 1, pp. 51–122, Jul. 2023, doi: 10.1007/s00454-022-00393-w.

[22] L. Cao and W. Ge, "Analysis of certificateless signcryption schemes and construction of a secure and efficient pairing-free one based on ECC," *KSII Transactions on Internet and Information Systems*, vol. 12, no. 9, pp. 4527–4547, Sep. 2018, doi: 10.3837/tiis.2018.09.022.

[23] Z. Šulc and H. Řezanková, "Comparison of Similarity Measures for Categorical Data in Hierarchical Clustering," *Journal of Classification*, vol. 36, no. 1, pp. 58–72, Apr. 2019, doi: 10.1007/s00357-019-09317-5.

[24] H. Zhou, Y. Zhang, and Y. Liu, "A Global-Relationship Dissimilarity Measure for the k -Modes Clustering Algorithm," *Computational Intelligence and Neuroscience*, vol. 2017, pp. 1–7, 2017, doi: 10.1155/2017/3691316.

[25] A. K. Das, S. Sengupta, and S. Bhattacharyya, "A group incremental feature selection for classification using rough set theory based genetic algorithm," *Applied Soft Computing*, vol. 65, pp. 400–411, Apr. 2018, doi: 10.1016/j.asoc.2018.01.040.

[26] A. R. Escobedo, E. Moreno-Centeno, and R. Yasmin, "An axiomatic distance methodology for aggregating multimodal evaluations," *Information Sciences*, vol. 590, pp. 322–345, Apr. 2022, doi: 10.1016/j.ins.2021.12.124.

[27] J. Xie, M. Wang, X. Lu, X. Liu, and P. W. Grant, "DP-k-modes: A self-tuning k -modes clustering algorithm," *Pattern Recognition Letters*, vol. 158, pp. 117–124, Jun. 2022, doi: 10.1016/j.patrec.2022.04.026.

[28] N. Hamidah and E. Istiyono, "The quality of test on National Examination of Natural science in the level of elementary school," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 11, no. 2, p. 604, Jun. 2022, doi: 10.11591/ijere.v11i2.22233.

[29] F. Farnoud Hassanzadeh and O. Milenkovic, "An Axiomatic Approach to Constructing Distances for Rank Comparison and Aggregation," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6417–6439, 2014, doi: 10.1109/TIT.2014.2345760.

[30] G. Sun, Z. Zhou, B. Chang, J. Tang, and R. Liang, "PermVizor: visual analysis of multivariate permutations," *Journal of Visualization*, vol. 22, no. 6, pp. 1225–1240, Dec. 2019, doi: 10.1007/s12650-019-00599-w.

[31] M. Mohandes, M. Deriche, and S. O. Aliyu, "Classifiers Combination Techniques: A Comprehensive Review," *IEEE Access*, vol. 6, pp. 19626–19639, 2018, doi: 10.1109/ACCESS.2018.2813079.

[32] M. Singla and K. K. Shukla, "Robust statistics-based support vector machine and its variants: a survey," *Neural Computing and Applications*, vol. 32, no. 15, pp. 11173–11194, Aug. 2020, doi: 10.1007/s00521-019-04627-6.

BIOGRAPHIES OF AUTHORS

Lienou T. Jean-Pierre obtained his MSc in System Engineering in Kiev Polytechnic Institute (National Technical University of Ukraine) and his PhD in University of Yaounde I (Cameroon). Former Maintenance Engineer at Labotech Medical, he was in charge of Medical imaging equipment. He joined the University of Dschang in 2002. He is actually Head of Department of Computer Engineering, College of Technology, of University of Bamenda. He is a member of IAENG since 2018. His Research fields are Method Engineering applied in control systems, Multi Agent Systems applied in power electric systems, cyber resilience and the use of various artificial intelligent techniques in diagnostic of complex systems. He manages a grant related to cyber resilience funded by US ARL. He can be contacted at email: lienou@gmail.com.

Djimeli-Tsajio Alain Bernard received B.S. (2001) and M.S. with thesis (2004) from the Faculty of Science of the University of Yaoundé 1 and the Ph.D. (2016) from the Faculty of Science of the University of Dschang, all in Cameroon in the field of Physics option Electronics. Since 2006, he has joined Fotso Victor University Institute of Technology of the University of Dschang as lecturer in the Department of Telecommunication and Network Engineering. He is member of UR.A.I.A of the present university where he is carrying out research in the field of Artificial Intelligence for biomedical process. He can be contacted at email: adjimeli@yahoo.fr.

Noulamo Thierry earned his Master's degree, Diploma of Advanced Studies (DEA), and Ph.D. in software engineering science from the University of Yaounde I in 2000, 2001, and 2010, respectively. He is currently working as lecturer in Department of Computer Engineering from Fotso Victor University Institute of Technology (IUT-FV) of Bandjoun, since 2004. He is a member of EANG since 2009, Life member of URAIA since 2008. He has published more than 11 research papers in reputed international journals and conferences. His main research work focuses on Process Automation using MDE and Muti-Agent approach. He has 18 years of teaching experience and 11 years of research experience. He can be contacted at email: thierry.noulamo@gmail.com.

Fotsing Talla Bernard obtained his Bachelor in Mathematics and Fundamental Computer Science in 1999 at the University of Dschang. He obtained his Master's degree, his Diploma of Advanced Studies and his Ph.D in Computer Science at the University of Yaoundé I. He is currently working as a Lecturer in the Department of Computer Engineering of the Fotso Victor University Institute of Technology (IUT-FV) of Bandjoun. He is a member of the Research Unit in Control systems and Applied Informatics (URAIA) of the IUTFV since 2010. His main research work focuses on the design of model-driven software architectures, the specification of formal languages for the description of models using formal tools such as Attributed Grammars, and very recently the use of algorithms of Machine Learning for the diagnosis of certain diseases. He has 13 years of teaching experience and a few years of research experience. He can be contacted at bfotsing@gmail.com.