Numeral Recognition Using Statistical Methods Comparison Study

Huda A. Rasheed
AL- Mustansiriyah University
College of sciences/ Department of mathematics

Nada A. Rasheed
Babylon University
School of Information Technology

Abstract

The area of character recognition has been received a considerable attention by researchers all over the world during the last three decades. However, in this research explores best sets of feature extraction techniques and studies the accuracy of well-known classifiers for Arabic numeral using the Statistical styles in two methods and comparison study between them. First methods Linear Discriminant function, that are yield results with accuracy as high as 90% of original grouped cases correctly classified. Second method we proposed algorithm, the results show the efficiency of the proposed algorithms, where it is found to achieve recognition accuracy of 92.9% and 91.4%. This is providing efficient is more than the first method.

Keywords: Numeral, Recognition, Discriminant

Introduction

Document image analysis systems can contribute tremendously to the advancement of the automation process and can improve the interaction between man and machine in many applications, including office automation, check verification and a large variety of banking, business and data entry applications [1].

Character recognition is a long-standing, fundamental problem in pattern recognition. It has been the subject of a considerable number of studies and serves many useful applications. Of the two major issues in character recognition, character shape representation and category assignment, classification has been, by and large, the subject of most studies. These studies assumed that descriptions of shape by basic characteristics such as curvature, tangents, and transform coefficients, are sufficiently expressive to justify focusing on classification. However, a good representation is as important as a good classifier for high performance [2].

There are two types of character recognition systems: on-line and off-line systems. Each system has its own algorithms and methods. The main difference between them is that in an on-line system the recognition is performed in the time of writing while the off-line recognition is performed after the writing is completed [3].

The recognition of hand4written numeral characters has been a topic widely studied during the recent decades because of both its theoretical value in pattern recognition and its numerous possible applications [4], one such area is the reading of courtesy amounts on bank checks. This application has been very popular in handwriting recognition research, due to the availability of relatively inexpensive CPU power, and the possibility to reduce considerably the manual effort involved in this task. Another application is the reading of postal zip codes in addresses written or typed on envelopes. The former is more difficult than the latter due to a number of differences in the nature of the handwritten material. For example, bank checks systems [5].

Pattern recognition systems typically involve two steps: feature extraction in which appropriate representation of pattern are developed and classification in which decision rules for separating pattern classes are defined. There are indeed as many possible features as the ways characters are written. These features can be classified into two major categories: statistical and

structural features. In the statistical approaches the input pattern is characterized by a set of N features and its description is achieved my means of a feature vector belonging to an N-dimensional space. On the other hand, in structural approaches it is assumed that the pattern to be recognized can be decomposed into simpler components (called primitive) and then described in terms of simple appropriate attributes of primitives and their topological relations [6].
This paper is structured as the following: Section 2 provides a brief overview of the liner discernment functions, Section 3 presents the feature extraction from handwritten numeral, Section 4 describes the overall experiments and results and Section 5 includes some discussion and comparison. Finally, Section 6 presents our recommendations.

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a method of finding such a linear combination of variables which best separates two or more classes. It aims at the classification of an object into one of *K* given classes based on information from a set of *p* predictor variables. Among the many available methods, the simplest and most popular approach is linear discriminant analysis (LDA)

Fisher Linear Discriminant Analysis means finds a linear transformation of predictor variables which provides a more accurate discrimination (see the right).

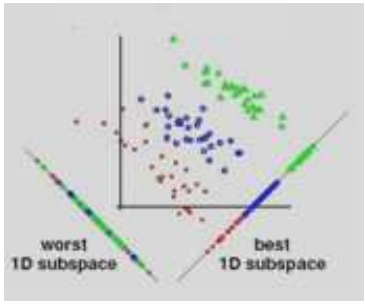

Figure1: class feature data.

● Objective is to find w to maximize J(w) , where:

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \quad \text{....} \quad (1)$$

● S_B is between class scatter matrixes, related to covariance matrix of centers of clusters

$$S_B = \sum N_c (\mu_c - \bar{x})(\mu_c - \bar{x})^T \quad \text{....} \quad (2)$$

●*Nc* is the number of observations in class c.

$$\mu_c = \frac{1}{N_c} \sum_{i \in c} x_i$$

$$\bar{x} = = \frac{1}{N} \sum_i x_i = \frac{1}{N} \sum_c N_c \mu_c$$

●S_W is within class scatter matrix, related to covariance matrix between clusters

$$S_W = \sum_c \sum_{i \in c} (x_i - \mu_c)(x_i - \mu_c)^T \quad \text{....} \quad (3)$$

Good Class Separation includes Finding the direction to project data on so that:
– Between classes variance is maximized.
– Within class Variance is minimized.

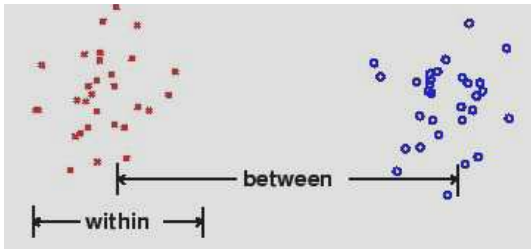

Figure2: Good class separation

Feature Extraction

The issue of how many features a method needs to use to obtain reliable handwritten numeral recognition is a very difficult one. There is always a temptation to include more and more features in a method in a hope of improving performance.

Most methods, but not all, during the feature extraction, would generate a reference numeral (or a set of reference numeral) for each individual. This normally requires a number of numeral of the user to be captured at enrollment or registration time (we call this numeral the sample numeral). When a user claims to be a particular individual and presents a numeral (we call this numeral the test numeral), the test numeral is compared with the reference numeral for that individual using one of the methods in this paper.

To preparing the numeral database is to determine the number of sample numeral that need to be used in recognition, here the sample is (70) digits, that collected from different persons and used their numeral in database.

To reduce the variation in size, all numeral images will be scaled at fixed size (77x77) pixels, if the scanned numeral is large or smaller, then it will be resized so that all numerals are having the same size. After the images were acquired, they were converted into monochrome bitmap (BMP) form, it was necessary to convert the images into binary representations of the handwriting.

In the present case, the numeral images were divided into 7 segments where each segment consists of 1×77 pixels, as in Figure (3), each segments is represented features or measurements as vector (length 0.0, length 0.2, length 0.4, length 0.5, length 0.6, length 0.8, and length 1.0). In addition to the value of tall feature that is represented the length column between the points (a) and (b) as shown in figure (3) below, this feature is necessary to different between (4 and 9) numerals.

It is viewed as a point in a dimensional space (8 x 70), the goal is to choose those features that allow pattern vectors belonging to different categories.

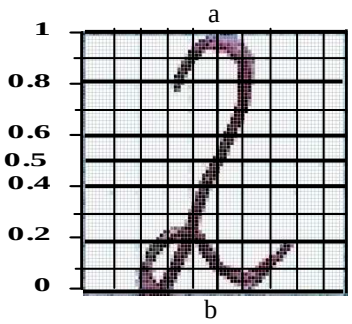

Figure3: Decomposition of a numeral (Features)

Materials and Methods

After feature extraction stage begins, the result of this stage is a vector of (8) values for each numeral and there are (70) numeral for different persons in all database.

1. Recognition using linear discriminant function

The efficiency of the linear discrimination technique has been tested over a wide range of handwritten character recognition problems.

We had several features extracted from the handwritten numeral, the vector (length.0, length.2, length.4, length.5, length.6, length.8, and length1, tall) for each numeral which created matrix dimension (8 x 70). We tried LDA using SPSS.v11 software (Statistical Package for Social Sciences) and applied Fisher's linear discriminant functions then we get the Classification Function Coefficients, results as shown in the table (1) below:

Table (1): Classification Function Coefficients by using Fisher's linear discriminant functions

LENGTH	GROUP									
	0	1	2	3	4	5	6	7	8	9
V1	.153	.215	.211	.416	.208	.760	.292	1.023	.461	.360
V.8	.307	-.014	.334	-.078	.344	-.248	-.087	-.073	.155	.305
V.6	.088	-.354	-.447	-.323	.409	-.312	-.720	-.619	-.666	-.107
V.5	.289	-.121	-.003	-.064	.155	-.016	.003	-.049	.136	-.075
V.4	.514	.205	.041	.197	-.087	.310	.358	.402	.454	.067
V.2	.105	-.450	-.357	-.379	-.403	-.129	.286	-.421	.230	-.497
V.0	.184	-.026	.220	-.008	-.034	-.010	.139	-.287	.122	-.100
TALL	3.536	3.996	3.928	3.954	1.756	3.525	3.778	4.104	3.762	3.912
(Constant)	-167.679	-154.851	-159.852	-154.126	-46.514	-133.942	-154.301	-178.802	-161.581	-156.848

To recognize numeral it must calculate the vector (length.0, length.2, length.4, length.5, length.6, length.8, length1, tall) to the unknown number. Then redress them in group of ten functions that computed according in the table (1) above for each numeral as shown below then the unknown numeral belong to the class which consideration the equation that afford from largest value.

The following table (2) summarizes the main outputs of LDA, which gives a good classification of 90% of original grouped cases correctly classified.

Group(0)= -167.679 + 3.536 (TALL) + .184 (LENGTH.0)+0.105 (LENGTH.2)+0.514 (LENGTH.4) + 0.289 (LENGTH.5) + 0.088(LENGTH.6) + 0.307(LENGTH.8) + 0.153(LENGTH1)

.
.
.

Group(9)= -156.848 + 3.912 (TALL) - .100 (LENGTH.0) - .497 (LENGTH.2) + .067 (LENGTH.4) - .075 (LENGTH.5) - .107 (LENGTH.6) + .305 (LENGTH.8) + .360 (LENGTH1)

Table (2): Classification Results according to Fisher Linear Discriminant

GROUP		Predicted Group Membership										Total
		0	1	2	3	4	5	6	7	8	9	
Count	0	7	0	0	0	0	0	0	0	0	0	7
	1	0	7	1	0	0	0	0	0	0	0	8
	2	0	0	7	0	0	0	0	0	0	0	7
	3	0	0	0	6	0	0	1	0	0	0	7
	4	0	0	0	0	8	0	0	0	0	0	8
	5	0	0	0	1	0	3	1	0	0	0	5
	6	0	0	0	0	0	0	6	0	2	0	8
	7	0	0	0	0	0	0	0	6	0	0	6
	8	0	0	0	0	0	0	0	0	7	0	7
	9	0	1	0	0	0	0	0	0	0	6	7
%	0	100.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	100.0
	1	.0	87.5	12.5	.0	.0	.0	.0	.0	.0	.0	100.0
	2	.0	.0	100.0	.0	.0	.0	.0	.0	.0	.0	100.0
	3	.0	.0	.0	85.7	.0	.0	14.3	.0	.0	.0	100.0
	4	.0	.0	.0	.0	100.0	.0	.0	.0	.0	.0	100.0
	5	.0	.0	.0	20.0	.0	60.0	20.0	.0	.0	.0	100.0
	6	.0	.0	.0	.0	.0	.0	75.0	.0	25.0	.0	100.0
	7	.0	.0	.0	.0	.0	.0	.0	100.0	.0	.0	100.0
	8	.0	.0	.0	.0	.0	.0	.0	.0	100.0	.0	100.0
	9	.0	14.3	.0	.0	.0	.0	.0	.0	.0	85.7	100.0

This presents successfully features, which are extraction and efficiently the method Fisher Linear Discriminant Analysis in recognize numeral.

2. Recognition using the proposed algorithm

After extracting feature vectors having (8) elements for each numeral available as the database, these vectors will represent the input to the numeral pattern recognizer. Then we are applying the proposed algorithm that can be presented as follows:

The proposed algorithm is shown as follows:

Step1: For i = 0 to 9
Step2: For j =1 to n_i , (n_i represented the number of replication of the class i, i=0,1,2,….,9)
Step3: Input features (Length.0,Length.2, Length.4, Length.5, Length.6, Length.8, Length1, Tall)
Step3.1: Next j
Step4: Depending on n_i values, calculate average to each feature: $\overline{Length.0i}, \overline{Length.2i}, \ldots\ldots, \overline{Talli}$ through class i , as a result it will be a vector of order 8.
Step5: Depending on n_i valus, calculate standard deviation to each feature:
S_{0i} , $S_{.2i}$, $S_{.4i}$, $S_{.5i}$, $S_{.6i}$, $S_{.8i}$, S_{1i} through class i , as a result it will be a vector of order 8.
Step5.1: Next i
Step6: For i = 0 to 9
Step7: Substitute features for unknown numeric (v0, v0.2, v0.4, v0.5, v0.6, v0.8, v1, v.Tall) into the following equation:

$$d_i = \frac{\left|v0 - \overline{Length.0i}\right|}{S_{oi}} + \frac{\left|v0.2 - \overline{Length.2i}\right|}{S_{2i}} + \frac{\left|v0.4 - \overline{Length.4i}\right|}{S_{4i}} + \frac{\left|v0.5 - \overline{Length.5i}\right|}{S_{5i}} + \frac{\left|v0.6 - \overline{Length.6i}\right|}{S_{6i}} + \frac{\left|v0.8 - \overline{Length.8i}\right|}{S_{8i}} + \frac{\left|v1 - \overline{Length.1i}\right|}{S_{1i}} + \left|v.Tall - \overline{Talli}\right| \quad (4)$$

Where i represents the class (i = 0,1,2,…,9) .
Step7.1: Next i
Step8: Classifying unknown number into class which corresponds to Minimum value of d_i

Firstly, the data used in recognition involves of the vectors values (Length.0, Length.2, Length.4, Length.5, Length.6, Length.8, Length1, Tall) average each features through each class to get the table (3) below:

Table (3): The Distribution of Features Averages According to its class

Class	Length1	Length.8	Length.6	Length.5	Length.4	Length.2	Length.0	Tall
0	15.43	35.43	41.57	42.43	43.00	39.00	16.86	77.00
1	3.75	7.63	3.88	4.00	4.13	4.38	12.13	77.00
2	10.00	25.57	3.43	3.71	4.14	6.00	25.14	77.00
3	13.43	6.57	9.00	9.14	4.57	8.43	13.71	77.00
4	18.13	25.00	26.75	20.13	7.63	4.88	3.88	36.88
5	29.20	4.00	16.40	15.40	10.40	18.60	13.40	70.00
6	6.50	7.88	6.25	7.63	12.63	33.38	16.38	77.00
7	38.83	18.67	5.83	12.33	15.67	5.17	4.33	77.00
8	19.57	26.86	13.43	19.43	27.29	32.57	17.29	77.00
9	17.86	29.00	16.57	4.86	5.14	9.86	7.43	77.00

Also, the standard deviations to each feature through classes 0, 1,…, 9 are shown in following table:

Table (4): The Distribution of Features Standard deviations According to its class

Class	Length1	Length.8	Length.6	Length.5	Length.4	Length.2	Length.0	Tall
0	7.96	6.02	6.13	6.80	6.86	6.30	7.73	.00
1	1.04	4.69	.64	.76	.99	.92	9.34	.00
2	3.11	7.50	.53	.76	.38	4.55	13.48	.00
3	4.31	7.30	5.72	6.15	1.40	8.75	9.98	.00
4	12.46	6.05	5.60	14.12	7.93	1.64	1.36	6.38
5	9.42	1.22	12.97	14.26	14.33	19.59	9.29	15.65
6	4.78	9.82	6.43	10.31	13.86	6.82	6.67	.00
7	4.45	16.17	.98	14.61	9.85	1.47	.52	.00
8	7.55	9.17	3.87	5.97	5.88	4.31	5.53	.00
9	7.99	4.76	12.08	1.68	2.54	10.33	2.94	.00

Now, using table (3) and (4) , we can classify any unknown numeral into its class easily by applying equation (4) depending on its features. Experimental results are then presented and analyzed 92.9% of original grouped cases correctly classified.

Table (5): Classification Results according to proposed method

GROUP		Predicted Group Membership										Total
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
Count	0	7	0	0	0	0	0	0	0	0	0	7
	1	0	7	0	1	0	0	0	0	0	0	8
	2	0	0	5	1	0	0	0	0	0	1	7
	3	0	0	0	6	0	0	0	0	0	1	7
	4	0	0	0	0	8	0	0	0	0	0	8
	5	0	0	0	0	0	4	1	0	0	0	5
	6	0	0	0	0	0	0	8	0	0	0	8
	7	0	0	0	0	0	0	0	6	0	0	6
	8	0	0	0	0	0	0	0	0	7	0	7
	9	0	0	0	0	0	0	0	0	0	7	7
%	0	100	0	0	0	0	0	0	0	0	0	100
	1	0	87.5	0	12.5	0	0	0	0	0	0	100
	2	0	0	71.4	14.3	0	0	0	0	0	14.3	100
	3	0	0	0	85.7	0	0	0	0	0	14.3	100
	4	0	0	0	0	100	0	0	0	0	0	100
	5	0	0	0	0	0	80	20	0	0	0	100
	6	0	0	0	0	0	0	100	0	0	0	100
	7	0	0	0	0	0	0	0	100	0	0	100
	8	0	0	0	0	0	0	0	0	100	0	100
	9	0	0	0	0	0	0	0	0	0	100	100

Example (1): Suppose that the features of unknown number where as follows:

Length1	Length.8	Length.6	Length.5	Length.4	Length.2	Length.0	Tall
40	6	35	5	4	42	14	77

Depending on means and Standard deviations of features for each class-as showing in tables 3,4 respectively- we applied Equation (4) and got these results:

class	0	1	2	3	4	5	6	7	8	9
di	21.08	126.37	82.63	15.73	58.03	10.15	14.17	76.15	19.72	15.01

Now, we can classify the unknown number into class 5 because the minimum value of di was 10.15 which correspond to class 5.

There is another proposed algorithm which is the same as the previous algorithm except replacing equation (4) in step 7 by the following equation:

$$d_i = \frac{\left(v0 - \overline{Length.0i}\right)^2}{S_{oi}} + \frac{\left(v0.2 - \overline{Length.2i}\right)^2}{S_{2i}} + \frac{\left(v0.4 - \overline{Length.4i}\right)^2}{S_{4i}} + \frac{\left(v0.5 - \overline{Length.5i}\right)^2}{S_{5i}} + \frac{\left(v0.6 - \overline{Length.6i}\right)^2}{S_{6i}} + \frac{\left(v0.8 - \overline{Length.8i}\right)^2}{S_{8i}} + \frac{\left(v1 - \overline{Length.1i}\right)^2}{S_{1i}} + \left|v.Tall - \overline{Talli}\right| \quad \ldots (5)$$

91.4% of original grouped cases correctly classified by using this method, as shown in table (6):

Table (6): Classification Results according to second proposed method

GROUP		Predicted Group Membership										Total
		0	1	2	3	4	5	6	7	8	9	
Count	0	7	0	0	0	0	0	0	0	0	0	7
	1	0	7	1	0	0	0	0	0	0	0	8
	2	0	0	7	0	0	0	0	0	0	0	7
	3	0	0	0	6	0	0	0	0	0	1	7
	4	0	0	0	0	6	0	0	0	0	2	8
	5	0	0	0	0	0	5	0	0	0	0	5
	6	0	0	0	0	0	0	7	0	1	0	8
	7	0	0	0	0	0	0	0	6	0	0	6
	8	0	0	0	0	0	0	0	0	7	0	7
	9	0	0	2	0	0	0	0	0	0	5	7
%	0	100.0	0.	0.	0.	0.	0.	0.	0.	0.	0.	100.0
	1	0.	87.5	12.5	0.	0.	0.	0.	0.	0.	0.	100.0
	2	0.	0.	100.0	0.	0.	0.	0.	0.	0.	0.	100.0
	3	0.	0.	0.	85.7	0.	0.	14.3	0.	0.	0.	100.0
	4	0.	0.	0.	0.	87.5	0.	0.	0.	0.	12.5	100.0
	5	0.	0.	0.	0.	0.	80.0	20.0	0.	0.	0.	100.0
	6	0.	0.	0.	0.	0.	0.	87.5	0.	12.5	0.	100.0
	7	0.	0.	0.	0.	0.	0.	0.	100.0	0.	0.	100.0
	8	0.	0.	0.	0.	0.	0.	0.	0.	100.0	0.	100.0
	9	0.	14.3	0.	0.	0.	0.	0.	0.	0.	85.7	100.0

Example (2): Suppose that the features of unknown number as in example (1), to classify unknown number into its group using the second proposed method:
Apply Equation (5) and get these results:

class	0	1	2	3	4	5	6	7	8	9
di	657	4317	2517	414	1079	**87.8**	381	1999	373	316

Now, we can classify the unknown number into class 5 because the minimum value of di was 87.8 which correspond to class 5.

Results and Discussion

The main conclusions of this paper can be summarized as follows:
From table (5), we can see (92.9%) of original grouped cases correctly classified by using the first proposed method, the second proposed method classified (91.4) of original grouped correctly, while the recognition using Fisher's linear discriminant functions method (using SPSS.V11) decreased into (90%) of original grouped cases correctly classified, as a result , we can arrange compared method according to its efficiency as follows:

1- The first proposed method (92.9%).
2- The second proposed method (91.4%).
3- Fisher's linear discriminant functions method (90%).

This namely the proposed algorithm can be depended in recognition the numerical.

Recommendations

The following recommendations for further research can be identified:

1- Employed the first proposed algorithm in pattern recognition and attempted the equation that be used in recognition (4) in form that increase the efficiency and accuracy the recognition.

2- Attempted to drawing the number as curves and fineding the equation to each numerical by used the partial derivation to be for Future Research.

References

1-Mostafa G. M., 2004,"An Adaptive Algorithm for the Automatic Segmentation of Printed Arabic Text", 17th The national conference for computer. Abu-al-Aziz king University. Arabia Saudi.:437-444.
المؤتمر الوطني السابع عشر للحاسب الآلي (المعلوماتية في خدمة ضيوف الرحمن)، جامعة الملك عبدالعزيز ، المدينة المنورة (صفر 1425 هـ / أبريل 2004م).

2- Ehsan N., N. Mezghani, A. Mitiche and R. de B. Johnston, 2006,"Online Persian/Arabic character recognition by polynomial representation and a Kohonen network", In Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006), 4: 683-686, Hong Kong, China, August

3- Ahmed M. Z. and M. S. Zakaria, 2004,"Challenges in Recognizing Arabic Characters", 17th the international conference for computer. Abu-al-Aziz king University. Arabia Saudi.:445-452.
المؤتمر الوطني السابع عشر للحاسب الآلي (المعلوماتية في خدمة ضيوف الرحمن)، جامعة الملك عبدالعزيز ، المدينة المنورة (صفر 1425 هـ / أبريل 2004م).

4- Suzete E. N. C , J. M. de Carvalho and R. Sabourin, 2002,"On the Performance of Wavelets for Handwritten Numerals Recognition", ICPR**,**Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) ,3**:** 127-130.

5- Luiz S. O., R. Sabourin, F. vio Bortolozzi and C. Y. Suen, 2002,"Automatic Recognition of Handwritten Numeral Strings: A Recognition and Verification Strategy", IEEE Transactions On Pattern Analysis And Machine Intelligence,24(11):1438-1454.

6- Saeed M., Karim F. and Majid Z., 2005,"A Hybrid Structural/Statistical Classifier for Handwritten Farsi/Arabic Numeral Recognition", In IAPR Conference on Machine Vision Applications (MVA), May 16-18, 2005 Tsukuba Science City, Japan, ISBN 4-901122-04-5 S.6-4:218-221.

تمييز الأرقام باستخدام الطرق الإحصائية
دراسة مقارنة

هدى عبدالله رشيد
الجامعة المستنصرية
كلية العلوم/قسم الرياضيات

ندى عبدالله رشيد
جامعة بابل
كلية التربية الأساسية

الخلاصـــــــة

مجال تمييز الأحرف يلقى عناية واسعة من قبل الباحثين في اكثر العالم خلال الثلاث العقود الأخيرة. على أية حال ، تم خلال البحث الكشف عن أفضل الخصائص المميزة بين الأرقام العربية باستخدام اسلوبين إحصائيين ومقارنة الدراسة بينهم.

أول طريقة هي الدالة الخطية المميزة ، حققت نتائج بدقة أعلى من 90% من المجموعة بتمييز الأرقام الأصلية . أما الطريقة الثانية فقد اقترحنا خوارزمية بأسلوبين ، النتائج تبين كفاءة الطريقة المقترحة التي وجدت للوصول إلى تمييز بدقة 92.9% و 91.4% وهذه تقدم كفاءة أكثر من الطريقة الأولى.