Iraqi Journal of Statistical Sciences (19) 2011
p.p.[39-48]

Eigen Values of Covariance Matrix for Feature Extraction of Latin Printed Image

khalil I. Alsaif* **Shaimaa M.Mohi Al-Deen****

Abstract

In this research the covariance matrix which was used in so many fields, its eigen values adopted to be the main parameters for Latin printed character recognition.

The idea was divided in two stages. The first stage is to generate the covariance matrix then, evaluate its eigen values to build main table for the whole latin characters in addition to the numeric values. The second stage is the recognition stage, which achieved to any character that was entered.

The applied examples do not register any negative result. So it can be strongly recommended for printed character recognition.

الملخص

في هذا البحث تم اعتماد القيمة المميزة لمصفوفة التباين (covariance) التي تستخدم في كثير من المجالات عاملا رئيسيا في تمييز الحرف اللاتيني المطبوع.

تقع الفكرة الرئيسية في مرحلتين، المرحلة الاولى لتكوين مصفوفة التباين لكل حرف من الاحرف المطبوعة ومن ثم حساب القيمة المميزة لها لاعتمادها في بناء قاعدة لجميع الحروف اللاتينية المطبوعة فضلا عن الاعداد، اما المرحلة الثانية فتمثل مرحلة التمييز لاي حرف لاتيني يتم ادخاله.

ان الامثلة المطبقة لم تسجل أية نتيجة سلبية، لذا يمكن التوصية باعتماد فكرة البحث لتمييز الحرف اللاتيني المطبوع.

* Asst. Prof./ Computer Sciences department/ College of Computer sciences & Mathematics/ University of Mosul/ IRAQ khalil_alsaif@hotmail.com

** Asst. Lecturer/ ./ Computer Sciences department/ College of Computer sciences & Mathematics/ University of Mosul/ IRAQ
shaima_mustafa76@yahoo.com

Received: 8/ 6 /2010 Accepted: 1 / 9 / 2010

1. Introduction:

Optical Character Recognition [OCR], is the process of converting the image obtained by scanning a text or a document into machine-editable format. OCR is one of the most important fields of pattern recognition and has been the center of attention for researchers in the last forty decades (Almohri, 2008).

Character recognition is one of the oldest fields of research. It is the art of automating both the process of reading and keyboard input of text in documents. A major part of information in documents is in the form of alphanumeric text (Zidouri, 2006).

The ultimate objective of any OCR system is to simulate the human reading capabilities. That is why OCR systems are considered a branch of an artificial intelligence and a branch of computer vision as well character recognition has received a lot of attention and success for latin languages (Amin, 2003).

The online problem is usually easier than the offline problem since more information is available. These two domains (offline & online) can be further divided into two areas according to the character itself that is either handwritten or printed character. Roughly, the OCR system is based on three main stages: preprocessing, feature extraction, and discrimination (called also, classifier, or recognition engine), (Fig. 1) depicts the block diagram of the typical OCR system (ABU RAS, 2007)(Liana, 2006).

(Figure-1) The typical OCR block diagram

In the review of the related work so many papers are published and beside the main goal of any OCR system which is simulating human's reading capability, the accuracy and time consuming are very important issues in this aspect. Based on the

latest survey which is published in May 2006 by Liana M. and Venu G. (Liana, 2006) all covered papers have presented their proposals seeking high accuracy and less time. Each one has treated the issue from different point of view. Their work can be classified into three main categories: preprocessing problems, features extraction problems, and recognition (discrimination) problems. For preprocessing stage, where the image is often converted to a more concise representation prior to recognition the most common methods are a skeleton which is a one-pixel thick representation showing the centre lines of the character (AL-Shatnawi, 2008).

*2. Concepts of the Covariance matrix (*Abdi, 2003)*:*

If X_1 and X_2 are strongly (positively) linked, random variables then we could think of defining covariance in a way that would embodies the following states:

* Whenever X_1 is positive, then X_2 is likely to be positive too.
* Whenever X_1 is negative, then X_2 is likely to be negative too.

This will not act because we want the covariance to be unchanged when both probability distributions are translated by arbitrary quantities. So instead of measuring the values of X_1 and X_2 from "0", we will measure them from reference points that translate along with the probability distributions, for example their respective means μ_1 and μ_2. Our original idea now reads:

* Whenever $(X_1 - \mu_1)$ is positive, then $(X_2 - \mu_2)$ is likely to be positive too.
* Whenever $(X_1 - \mu_1)$ is negative, then $(X_2 - \mu_2)$ is likely to be negative too.

So if X_1 and X_2 are strongly (positively) linked, more often than not, $X_1 - \mu_1$ and $X_2 - \mu_2$ are :

* Simultaneously positive,
* Or simultaneously negative.

The product $(X_1 - \mu_1).(X_2 - \mu_2)$ is then likely to be very often positive :

* Either because both quantities are positive,
* Or because both quantities are negative.

Yet, the product $(X_1 - \mu_1).(X_2 - \mu_2)$ is a random variable, and we want a fixed number. But a random variable that spends most of its time taking positive values is likely to have a positive expectation. So we will consider the **expectation** of $(X_1 - \mu_1).(X_2 - \mu_2)$, and call it the **covariance** of X_1 and X_2 as shown in equation (1).

$$Cov(X_1, X_2) = E[(X_1 - \mu_1).(X_2 - \mu_2)] \quad (1)$$

We'll show that this expression is equivalent to the other one, more convenient in practice can be seen clearly in equation (2).

$$Cov(X, Y) = E[XY] - E[X].E[Y] \quad (2)$$

3. Properties of a covariance matrix (Harris,2001)(Strang, 2003):

The covariance matrix is not just a convenient way of displaying numbers. As a matrix, it has several important properties which derive from the fact that a covariance matrix is always positive semi definite. The converse is also true: any positive semi definite matrix $\Sigma$is the covariance matrix of a random vector (in fact, of many).

In particular, the spectral decomposition of the covariance matrix of a random vector ***x*** shows that:

* There exists an orthonormal basis such that the covariance matrix Σ of ***x*** expressed in this basis is **diagonal**. The axes of this new basis are called the **principal components** of Σ(or of the distribution of ***x***).

* As the off-diagonal elements of this new matrix are 0, the new variables defined by this new basis (the projections of ***x*** on the principal components) are **uncorrelated**.

* The diagonal elements of this new, diagonal covariance matrix are the **Eigen values** of Σ. So the variances of the projections of ***x*** on the Principal Components are equal to the corresponding Eigen values of Σ.

* If units are changed so that all Principal Components now carry the same variance, the distribution is said to be "sphericized" (which is an abuse of language as the distribution is not necessarily spherically symmetric) : the marginal variables are now standardized **and** uncorrelated.

1. Construct an average signal M as shown in equation (3).

$$M_j = \frac{1}{N}\sum_{i=0}^{N-1} X_{[i][j]} \qquad j = 0.... L-1 \qquad (3)$$

2. Subtract the average signal from the original ensemble

$$Y_{[i]} = X_{[i]} - M \qquad i = 0.... N-1$$

3. Construct a covariance matrix.

$$\begin{pmatrix} C_{1,1} & C_{1,2} & \cdots & \cdots & C_{1,L} \\ C_{2,1} & C_{2,2} & & & \\ \vdots & \vdots & \ddots & & \\ \vdots & \vdots & & \ddots & \\ C_{L,1} & C_{L,2} & & & C_{L,L} \end{pmatrix}$$

Where each value $C_{i,j}$ is given in equation (4).

$$C_{i,j} = \sum_{p=0}^{N-1} \frac{Y_{[p][i]} \times Y_{[p][j]}}{N-1} \qquad i = 1...L,\ j = 1...L \qquad (4)$$

4. Proposed Algorithm:

In this paper an approach for printed character recognition is suggested and to be studied by applying different printed characters for recognition (all applied examples done on standard fonts), the procedures of the proposed algorithm can be classified in two stages:

a) Stages of preparing the database table:

1. Image acquisition from any input devices or via stored files on hard disk.

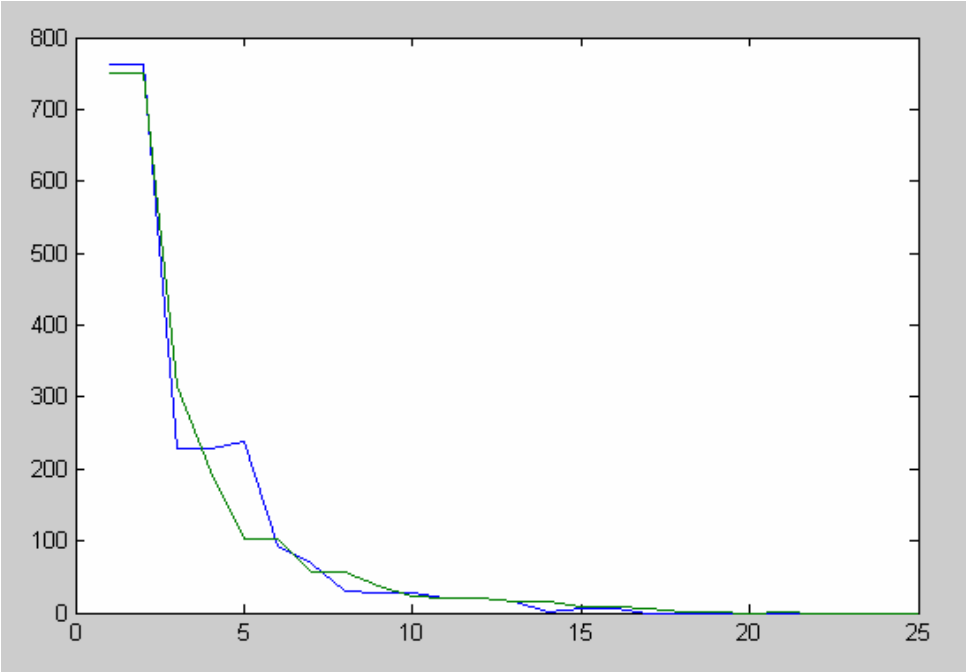
2. Preprocess to be done on the character image and resize it to 64×64 pixel.
3. Create the covariance matrix of the image character.
4. Evaluate the Eigen value of the covariance matrix.
5. Store the Eigen value on database table (in addition to the character name).
6. Step 1-5 to be run over all Latin characters plus the numeric values.

b) Stages of Recognition:

1. Get any printed Latin character.
2. Preprocess to be done in addition to resize it to 64×64.
3. Covariance matrix to be prepared in the same way of stage 1.
4. Evaluate the eigen values of the covariance matrix.
5. Look for the nearest set (of the evaluated eigen value) to the eigen values which were stored in the database table.
6. If does not find, the new set of eigen values to be added to the table, otherwise the character was recognized.

5. Results discussion:

When applying the proposed algorithm on closed characters such as: E& F, b& d and O& Q the eigen values for each pair [closed characters] shows the clear difference. In table (1) and Fig. 2 clear difference can be seen which can be adopted for recognition between the other characters. 25 Eigen values were used in the practical but the first 15 Eigen values could be enough to give good recognition between the characters.

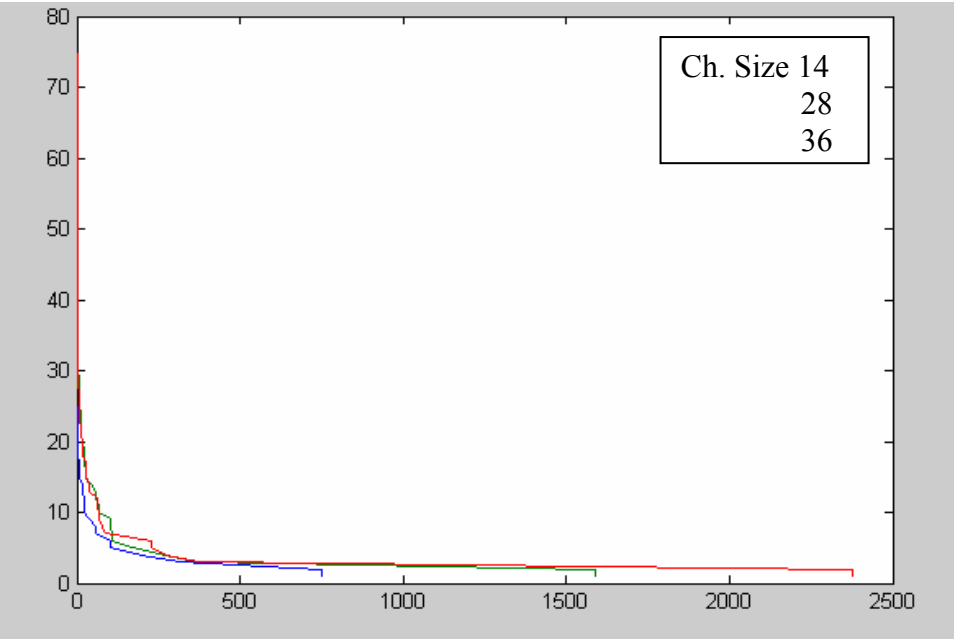


(Figure-2) the difference between character E and F

Table (1) The eign values of the characters

A	495.672631	158.573279	146.8987	123.820204	123.820204	58.107667	54.092931	41.171547	2.662942	7.424035	7.424035	7.661362
B	1050.613480	164.431056	197.3289	134.567439	23.865256	26.513256	28.503150	28.503150	22.664176	10.035364	4.926073	0.152360
.												
.												
y	552.259118	320.855971	245.2524	107.366231	57.074155	46.483631	32.513676	23.939316	23.939316	19.668711	14.391300	12.044999
Z	520.294708	198.820150	202.5428	141.778121	49.658491	39.982060	39.313920	39.313920	25.298985	25.298985	21.662273	5.360285
a	495.672631	158.573279	146.8987	123.820204	58.107667	54.092931	41.171547	41.171547	2.662942	7.424035	7.424035	7.661362
b	1050.613480	164.431056	197.3289	134.567439	23.865256	26.513256	28.503150	22.664176	10.035364	4.926073	4.926073	0.152360
.												
z	520.294708	198.820150	202.5428	141.778121	49.658491	39.982060	39.313920	39.313920	25.298985	25.298985	21.662273	5.360285
0	765.929239	153.003309	43.87356	27.057177	20.270030	13.849998	6.981071	6.667279	6.082227	2.034915	0.478420	0.079458
1	495.672631	158.573279	146.8987	123.820204	58.107667	54.092931	54.092931	41.171547	2.662942	7.424035	7.424035	7.661362
.												
9	1010.705994	1010.705999	187.8087	15.108846	15.108846	6.886475	4.864172	2.174813	2.174813	0.000000	0.000000	0.000000

Also different sizes of same character were studied to see the effect of the character size. (Fig. 3) show the eigen value of character [E] with wide range of size.



(Figure-3) character E in different sizes

To get standardization for the proposed algorithm far away from the problems of character size, a fixed size for different character size was adopted, i.e. the proposed algorithm arranges the character size to be of [64×64] with this suggestion, the result goes for high rate of recognition.

In the appendix, the algorithm tries to cover the most closed characters and to show clear difference in the values of their Eigen values. The curves given in the appendix for character Q& O can be adopted as reasonable to recommend the proposed algorithm for character recognition.

Conclusion and future work:

The proposed algorithm which was tested on most of the printed Latin characters does not register any negative result.

The applied algorithm can be developed by adding Neural Networks to work as automatic classifier, also can be used to evaluate the value of numeric image.

References:

[1] Haidar Almohri, (2008), "A Real-time DSP-Based Optical Character Recognition System for Isolated Arabic characters using the TI TMS320C6416T", Proceedings of The 2008 IAJC-IJME International Conference, ISBN 978-1-60643-379-9,.

[2] Abdelmalek Zidouri, (2006), "ORAN SYSTEM: A BASIS FOR AN ARABIC OCR", The Arabian Journal for Science and Engineering, Volume 31, Number 1B.

[3] Amin A., (2003),"Recognition of Hand-Printed Characters Based on Structural Description and Inductive Logic Programming", Pattern Recognition Letters, 24, pp. 3187-3196.

[4] Abdurazzag Ali ABU RAS and Salem M.A.REHIEL, (2007), "Off-line Omni-style Handwriting Arabic Character Recognition System Based on Wavelet Compression", ARISER, Vol. 3, No.4, pp.123-135.

[5] Liana M& Venu G., (2006), "Off line Arabic Hand writing Recognition: A Survey", IEEE, Transactions On Pattern Analysis and Machine Intelligence, 28(5), pp.712-724.

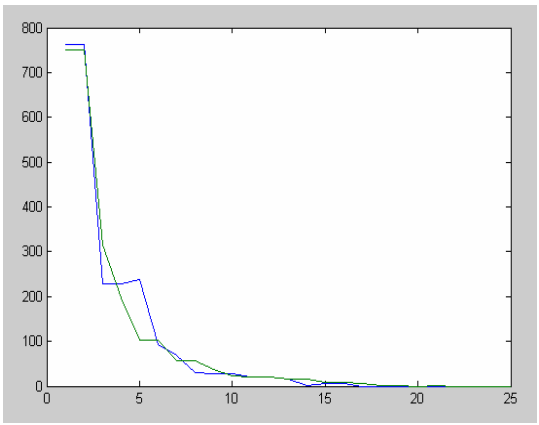
[6] Atallah AL-Shatnawi and Khairuddin Omar, (2008), " Methods of Arabic Language Baseline Detection – The State of Art", IJCSNS International Journal of Computer Science and Network Security, VOL.8, No.10.

[7] Abdi, H., (2003), "Multivariate analysis", In M. Lewis-Beck, A. Bryman, & T. Futing (Eds): Encyclopedia for research methods for the social sciences.

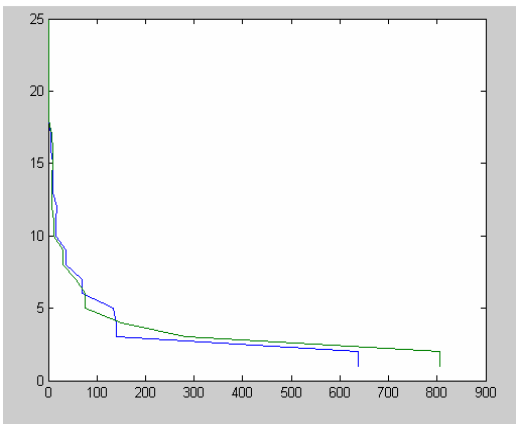
[8] Harris, R. j., (2001), "A primer of multivariate statistics", Mahwah (NJ): Lawrence Erlbaum, Associate publisher.

[9] Strang, G., (2003), "Introduction to linear algebra", Cambridge (MA): Wellesley- Cambridge Press.

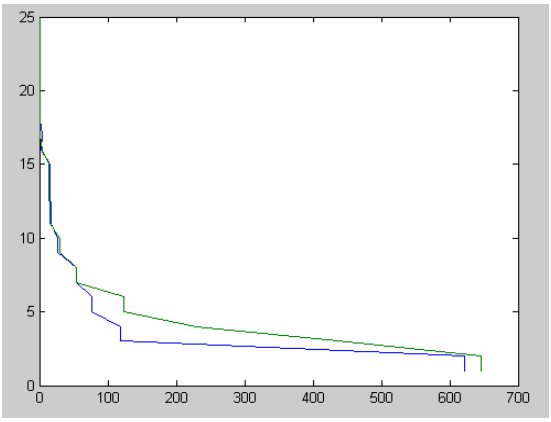
Appendix
Tested character (E& F), (d& b), (O&Q) and (I& l)



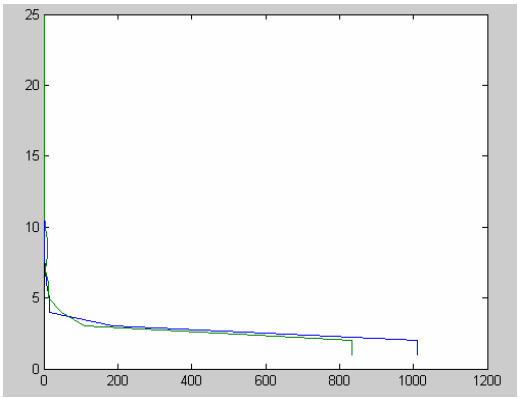
Character E and F



Character d and b



Character O and Q



Character I and l