AN INVESTIGATION ON EFFICIENT FEATURE EXTRACTION APPROACHES FOR ARABIC LETTER RECOGNITION

H. Aboaisha[1], Zhijie Xu[1], I. El-Feghi[2]
h.aboaisha@hud.ac.uk, z.xu@hud.ac.uk, Idrisel@ee.edu.ly
[1]School of Computing and Engineering, University of Huddersfield, Queensgate, Huddersfield, UK, HD1 3DH
[2]Electronic Engineering Department, University of Tripoli, Tripoli, Libya

ABSTRACT

Invariant features play an essential role in many pattern recognition applications due to their robustness to different environmental settings. In this research, two of the invariant moments, the Zernike Moment (ZM) and the Legendre Moment (LM), have been investigated for their suitability and computational efficiency in Arabic letter recognition. This paper starts with an introduction on Arabic letter characteristics before moving on to a literature review of current letter recognition strategies and systems. The concepts and algorithms of the ZM and LM techniques are then examined. To validate the new approach, a prototype system has been developed with several experiments carried out using a standard database called *IF/ENIT* which contains handwritten Tunisian town names and this database is used by many research groups working on recognition systems.

Keywords: Zernike Moments (ZM), Legendre Moments (LM), Optical Character Recognition (OCR), Arabic Letter Recognition

1. INTRODUCTION

In the last two decades, a significant amount of research work has been carried out in image processing and pattern recognition with applications being reported in face recognition, fingerprint identification, and letter recognition (Cheriet 2007). Although progress has been made on many fronts, efficient and effective Arabic letter recognition is still an active and challenging research field due to its intrinsic and difficult nature. Arabic letters can be found in more than 27 languages such as Arabic, Urdu, Jawi (Malay), Kurdish and Farsi (Persian) and are used by more than 260 million people across the globe (Cheung, Bennamoun et al. 1998). However, it has only recently been mentioned (Lorigo and Govindaraju 2006) that researchers have realised the difference of the Arabic letters from other popular alphabetic systems, such as English and other Latin-based letters. The most distinctive characteristics of Arabic letters are their cursive shapes of handwritten style and the complex combinatorial rules. In order to implement an Arabic letter recognition system, the selection rule for the feature vectors (or signatures) is of vital importance. The challenges for recognising a hand-written Arabic letter can be summarised as follows:

- Letters and words are extremely cursive and often similar in shape;
- The same letter at the beginning and end of a word can have a different form completely;
- Words can contain dots to change meanings;
- Some Arabic words can be divided into sub words in the so-called Piece of Arabic Words (PAWs) forms.

The so-called "Moment" features are basically scalar values being used to classify a function and to extract its meaningful features. In this research, the selection of the Moments as feature extractors was rooted from the fact that they have been proven effective as a shape feature representation scheme for complex contours. For example, trademarks and logos that are often over-complicated to be described by a single contour shape for relevant similarity measurements (Yong-Sung and Whoi-Yul 1997).

The rest of the paper is organised as follows: Section **2** introduces the invariant feature extraction techniques and the classic Optical Character Recognition (OCR) approaches. Section **3** highlights various invariant moments and their characteristics, including the Zernike Moments (ZM) and the Legendre Moments (LM). A system prototype for evaluating their performance in extracting Arabic letter features has been constructed. The experiment design and results are discussed in section **4**. Section **5** concludes the progress-to-date.

2. CHARACTER FEATURE EXTRACTION APPROACHES

The main aim of feature extraction is to establish appropriate and suitable feature vector signatures that encapsulate the most meaningful and important information about the examined objects. It is one of the core processes in all pattern recognition systems. To achieve a successful detection or retrieval task in digital image processing (DIP) systems, the feature selection rules need to be justified with great care (El-Feghi, Tahar et al. 2011). Any digital images can be represented by a collection of pixels. Thus, for classifying and recognising the properties of individual or group of pixels feature “descriptors” are often formed, which are usually just a set of numbers in the predefined formats and ranges. From this point of view, the studied subject patterns can be compared and recognised by matching the descriptors of an object pattern against the extracted descriptors from unknown objects. For evaluating the recognition results, effective descriptors should have four properties. Firstly, they should be able to determine a complete feature set. Secondly, they should be congruent, and capable of recognising similar pattern descriptors. Thirdly, they should be predominately based on invariant properties to ensure robustness. For instance, rotation-invariant features can be useful for recognising shapes independent from their orientational changes. Finally, they should be tolerant to the affine changes. Based on the literature review carried out in this programme, most of today’s recognition systems only take into account the subsets of these four criteria (Nixon and Aguado 2008).

Generally speaking, there are two types of feature extraction approaches, the low-level feature strategies and the high-level feature. The low-level features are well studied and can be extracted automatically from an image through statistical means such as colour histogram or lines and corners. The high-level feature methods are concerned with finding “shapes” in spatial terms that are often heavily influenced by pattern size and orientations.

2.1 LOW-LEVEL FEATURES AND THEIR EXTRACTION

There are several low-level feature based techniques being reported in extracting letters.

- **First-order edge detection**

The first-order edge detection method is based on a differentiation operation that denotes changes in image intensity. It can highlight image variations and contrast detecting boundaries of features within an image. The first-order algorithms are often used in edge-detection including Roberts Cross, Sobel and Canny filters. For example, the Roberts Cross filter seeks the boundary on the principle of gradient changes that can be obtained through any pairs of orthognality differences. This operator has the high accuracy and valuable results in vertical and horizontal positioning but is sensitive to noise.

The Sobel differential operator contains two masks to identify an edge in the vector form. Compared to other edge detecting operators, Sobel has two distinctive advantages (Wenshuo, Xiaoguang et al. 2010). Firstly, the extracted edges are explicit two columns/rows based differentiation; secondly, it has an inherent smoothing effect and is more robust to random noise. The Canny algorithm is an optimal edge detector based on a specific mathematical model involving four essential processes (Umbaugh 2008). It starts to smooth an input image by applying a Gaussian filter; then the filtering window scans direction and magnitude of the gradient. The nonmaxima suppression then follows to thinning the marked edges before finally a low threshold is applied to connect edge points denoted pixels.

- **Second-order edge detection**

Lablacian edge detection (Umbaugh 2008) is based on a template that creates second order differential results using two adjacent first-order differences as shown in eq.1:

$$f''(x) \cong \acute{f}(x) - \acute{f}(x+1) \tag{1}$$

Marr and Hildreth (1980) combined Gaussian smoothing with the second-order differentiation. The computation of the Laplacian and Gaussian can be approximated by convolving two Gaussian filters with different variance as shown in eq.2; where $g(x, y, \sigma)$ is the Gaussian function and k is a constant.

$$\sigma\nabla^2 g(x,y,\sigma) = \frac{\partial g}{\partial \sigma} \approx \frac{g(x,y,k\sigma) - g(x,y,\sigma)}{k\sigma - \sigma} \tag{2}$$

2.2 HIGH-LEVEL FEATURES AND THEIR EXTRACTION

In high-level feature extraction methods, the search for invariance properties is the paramagnet task. It steadily produces extraction results that are resistant to setting changes but focuses on the “contents” in an image. The four basic invariances studied in this project include illumination level invariance; position, location or translation invariance; rotation or orientation invariance; size or scale invariance.

- **Region descriptor**

A region describes interior points which are surrounded by a perimeter or boundaries. To describe the areas of a shape, there are basic measuring terms such as area, perimeter, compactness, etc. Also, there are more advanced shape description terms such as “Moments” which are global description tools. They are akin to combining compactness, area and other higher order descriptors together. Moments are associated more with statistical pattern recognition than with model-based vision, since a major assumption is that there is an unclouded view of the target shape (Nixon and Aguado 2008).

2.3 OPTICAL CHARACTER RECOGNITION

Optical Character Recognition (OCR) techniques are widely used in natural language processing, archiving and documentary management. In online recognition, the handwriting is generated by means of an electronic pen or a mouse for acquiring time-dependent signals (Liwicki and Bunke 2007), while in the offline mode, the text to be recognised is captured by a scanner and stored as an image. The difference between the online and offline handwriting recognition systems is depicted in Fig.1. Processes and issues relating to a recognition system include pattern representation, definition of pattern classes, sensing environment, feature extraction and selection, cluster analysis, selection of training and test samples, classifier design and learning, and assessment and evaluation (Jain, Duin et al. 2000). In dealing with classification problems, a classifier is often estimated to support some desired invariant properties. For instance, the shifting invariance in character recognition is important so the changes of a character location should not influence its clustering or classification results. A new approach has been introduced (Menasri, Vincent et al. 2007) based on the explicit grapheme segmentation method which incorporates the Hybrid Hidden Markov (HMM) model with a neural network-based recognition scheme. The advantages of this approach are rooted to its exploration of the redundancy in the shapes of Arabic letters.

3. INVARIANT MOMENTS

Recognising the objects and patterns which are deformed in many ways have been a research hot-spot in the last decade. There are three basic techniques applied in the field, namely, Brute Force, Image Normalization and Invariant Features. The invariant features describe the objects by a set of measurable scatter quantities. Invariant means it is insensitive to deformation and is capable of providing sufficient discrimination power to differentiate objects in different classes (Flusser, Zitova et al. 2009). Moments are numerical quantities that can be used to designate a function and also to extract meaningful features. There are popular moments techniques such as Zernike Moments, Hu Moments and Legendre Moments. Zernike Moments polynomials were first introduced in 1934 by Zernike. Hu has derived his seven common invariants of aspects (Ming-Kuei 1962).

3.1 ZERNIKE MOMENTS

ZM is considered to be one of the most powerful techniques in deploying orthogonal moments. The main justifications for introducing orthogonal moments are due to their fast and stable numerical implementations. The notion of orthogonal moments to extract invariant features from scalar quantities (moments) depends on the theory polynomials highlighted by Teague (Teague 1980). Theoretically, the polynomial bases generate the same space of functions because they are of the same degree. However, when applied, issues often raise computational efficiency and performance stability aspects. The image reconstruction from orthogonal moments can be fulfilled relatively easily. The reconstruction is optimal due to the orthogonal moments reducing the mean-square error.

The ZM polynomials are defined as follows:

$R_{nl}(r)$ is a real-valued Zernike-radial polynomial defined as follows:

$$R_{nl}(r) = \sum_{s=0}^{(n-|l|)/2} (-1)^s \frac{(n-s)!}{s!\left(\frac{n+|l|}{2}-s\right)!\left(\frac{n-|l|}{2}-s\right)!} r^{n-2s} \quad (3)$$

Here $n = 0,1,2,\ldots$, $l = -n, -n+2, \ldots, n$ and order of the polynomial is denoted by n and the repetition by l.

$= \sum_{k=|l|,|l|+2,...}^{n} B_{nlk}\, r^{k}$, the coefficients $B_{nlk}= \frac{(-1)^{(n-k)/2}((n+k)/2)!}{((n-k)\backslash 2)!((k+l)\backslash 2)!((k-l)\backslash 2)!}\, r^{n-2s}$

Where the B is the coefficient and l, k are the order of repetition.

$$Z_{nm} = \frac{n+1}{\pi} \sum_{r}\sum_{\theta} f(r\cos\theta, r\sin\theta).\, R_{nm}(r).\exp(jm\theta) \quad (4)$$

Where n is the radial magnitude, m is the radial direction and θ is the angle between the vector r and the x axis. There are many ways to convert from Cartesian coordinates to polar coordinates according to the particular applications. Commonly, it is supposed that the centre of the rotation is at the centroid of the image. The r represents the length of the vector from the origin to the (x, y) pixel. If image has size **H**eight ×**W**idth

$$r = \sqrt{(2*X-H+1)^2 + (2*y-w+1)^2} \;/H \quad (5)$$

$$\theta = arctan\left(\frac{W-1-2*y}{2*X-H+1}\right) \quad (6)$$

3.2 LEGENDRE MOMENTS

Legendre Moments belong to the class of orthogonal moments on rectangle. The main advantage of the LM is that it keeps the orthogonality even on the sampled image (Flusser, Zitova et al. 2009). The Legendre Moments of order (n, m), with image intensity function $f(x, y)$, is defined as:

$$L_{mn} = \frac{(2m+1)(2n+1)}{4MN} \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} P_{m}\left(-1+2\frac{x}{M-1}\right) P_{n}\left(-1+2\frac{y}{N-1}\right) \quad (7)$$

Where x, y are the pixel location. The Legendre polynomials are a complete orthogonal basis set defined over the interval [-1,1]. For example: if the m=5, n=4 and we have image size $M \times N$=3×3.

Representation of the pixel of the image $\begin{bmatrix} f(0,0) & f(0,1) & f(0,2) \\ f(1,0) & f(1,1) & f(1,2) \\ f(2,0) & f(2,1) & f(2,2) \end{bmatrix}$ and to compute $P_n(x)$ by Eq.8.

$$P_n(x) = \sum_{p=0}^{n} a_{np}\, x^{p} \quad (8)$$

If we are denoting the coefficients of the legend polynomials as a_{np}, the coefficients a_{np} can be defined as:

$$a_{np} = \begin{cases} (-1)^{(n-p)/2} \frac{1}{2^n} \frac{(n+p)!}{((n-p)/2)!((n+p)/2)!p!} & n-p\ even \\ 0 & n-p\ odd \end{cases} \quad (9)$$

3.3 SYSTEM PROTOTYPING AND PROCESS PARADIGM

The proposed system prototype comprises two processing stages: firstly, template Arabic letter features are extracted using ZM and stored in a database. The following phase will be the similarity measured between the input Arabic letters and the entire feature database to recognise any matched letters. Table 1 illustrates the Arabic letters tested in the project. The process pipeline of the proposed system is shown in Fig.2.

- **Zernike Moment Algorithm:**

To transfer ZM equation to a program as follows:
- Step 1. Compute the real-valued polynomial or radial;
- Step 2. Compute the Theta;
- Step 3. Compute r, the length of the vector from the origin to the (x, y);
- Step 4. Compute the Zernike Moments polynomials.

4. PRELIMINARY EXPERIMENTAL RESULTS

In the experiment design, focus was given to the comparison between the ZM and LM descriptors when applied to Arabic letters. The results demonstrated that the proposed system improved the precision-and-recall rate considerably from traditional low-level feature based methods. The developed Arabic recognition approaches using both ZM and LM have been evaluated and tested on the *IFN/ENIT* database (El Abed and Margner 2007), which is a standard test base for many Arabic character recognition systems containing more than 2,200 binary images of handwriting samples from over 400 writers. Fig.3 illustrates examples from the *IFN/ENIT* database for a Tunisian village name written by 12 different writers. The feature vector for ZM represents each word image in the database and each signature vector contains 19 dimensions, whereas in the LM each word image is represented by a signature vector of 45 dimensions. The ZM approach was able to recognise 70% of the word correctly even when the image is rotated. The LM technique was able to identify words with 60% accuracy. For example to search this word "طارث ", the ZM system recognises 7 out of 11 samples and the LM technique managed 5.

5. CONCLUSIONS

In this paper, Zernike and Legendre Moments for Arabic letter recognition have been investigated. Experiments demonstrated both methods' effectiveness in extracting and preserving Arabic letter characteristics. ZM is used due to its ability to compute the complex orthogonal moments precisely. The system has achieved satisfactory performance when compared with other OCR systems. The translational and scaling invariant, on the other hand, had struggled in LM to detect rotational invariant forms in the experiments. The project's objective for maximising the correct matching and retrieval from the Arabic database while minimising the false positive rate has been achieved.

REFERENCES

Cheriet, M. (2007). Strategies for visual arabic handwriting recognition: Issues and case study. Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on.

Cheung, A., M. Bennamoun, et al. (1998). A recognition-based Arabic optical character recognition system. Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on.

El-Feghi, I., A. Tahar, et al. (2011). Efficient features extraction for fingerprint classification with multi layer perceptron neural network. Systems, Signals and Devices (SSD), 2011 8th International Multi-Conference on.

El Abed, H. and V. Margner (2007). The IFN/ENIT-database - a tool to develop Arabic handwriting recognition systems. Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on.

Flusser, J., B. Zitova, et al. (2009). Moments and moment Invariants in Pattern Recongnition. Hoboken , USA, Wiley.

Jain, A. K., R. P. W. Duin, et al. (2000). "Statistical pattern recognition: a review." Pattern Analysis and Machine Intelligence, IEEE Transactions on 22(1): 4-37.

Liwicki, M. and H. Bunke (2007). Combining On-Line and Off-Line Systems for Handwriting Recognition. Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on.

Lorigo, L. M. and V. Govindaraju (2006). "Offline Arabic handwriting recognition: a survey." Pattern Analysis and Machine Intelligence, IEEE Transactions on 28(5): 712-724.

Menasri, F., N. Vincent, et al. (2007). Shape-Based Alphabet for Off-line Arabic Handwriting Recognition. Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on.

Ming-Kuei, H. (1962). "Visual pattern recognition by moment invariants." Information Theory, IRE Transactions on 8(2): 179-187.

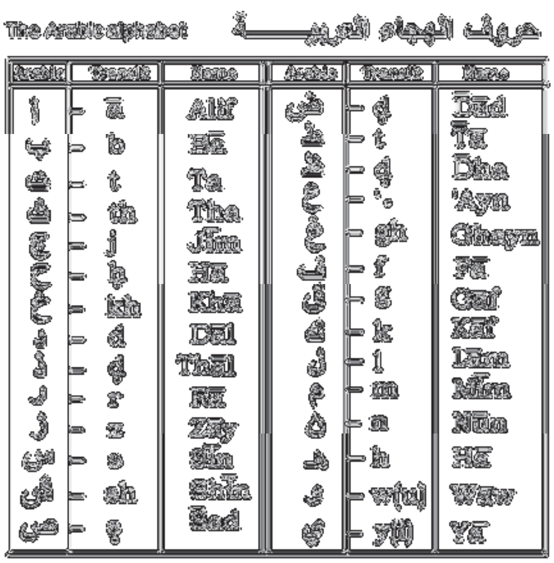
Nixon, M. and A. Aguado (2008). Feature Extraction & Image Processing, Academic Press.

Teague, M. R. (1980). "Image analysis via the general theory of moments*." J. Opt. Soc. Am. 70(8): 920-930.

Umbaugh, S. (2008). Computer Imaging: Digital Image Analysis and Processing, CRC Press.

Wenshuo, G., Z. Xiaoguang, et al. (2010). An improved Sobel edge detection. Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on.

Yong-Sung, K. and K. Whoi-Yul (1997). Content-based trademark retrieval system using visually salient features. Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on.


The Arabic alphabet حروف الهجاء العربية

Arabic	Translit	Name	Arabic	Translit	Name
ا	a	Alif	ض	d	Dad
ب	b	Ba	ط	t	Ta
ت	t	Ta	ظ	d	Dha
ث	th	Tha	ع	'	'Ayn
ج	j	Jim	غ	gh	Ghayn
ح	h	Ha	ف	f	Fa
خ	kh	Kha	ق	q	Qaf
د	d	Dal	ك	k	Kaf
ذ	d	Thal	ل	l	Lam
ر	r	Ra	م	m	Mim
ز	z	Zay	ن	n	Nun
س	s	Sin	ه	h	Ha
ش	sh	Shin	و	w(u)	Waw
ص	s	Sad	ي	y(i)	Ya

Table1. Arabic letters

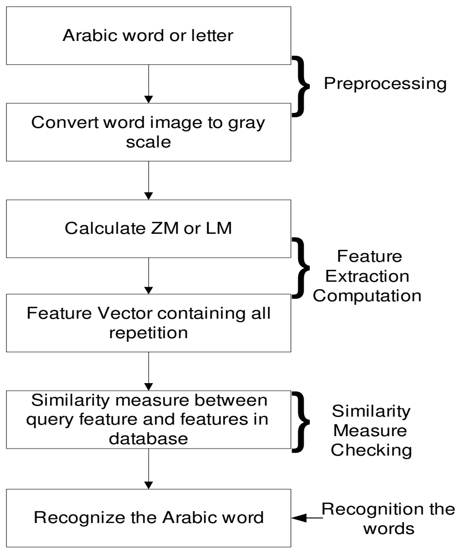


Figure2. Recognition system stages

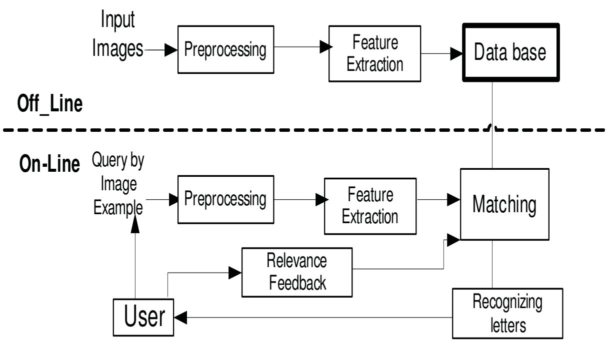


Figure1. The difference between on-line and off-line handwriting recognition

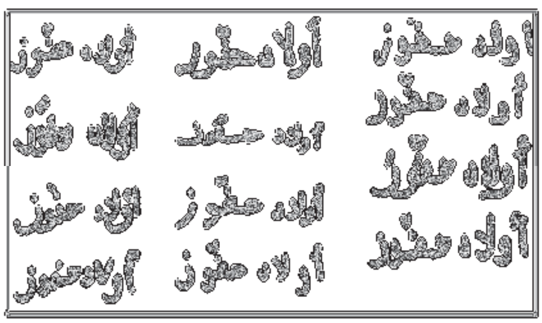


Figure3. Samples from *IFN/ENIT* database for Tunisian village name written in different styles.