Swarm-Based Feature Selection for Handwriting Identification

[1,2]Khaled Mohammed Bin Abdl and [2]Siti Zaiton Mohd Hashim
[1]Faculty of Applied Science, Hadramout University for Science and Technology, Yemen
[2]Faculty of Computer Science and Information Systems, University Technology Malaysia, Malaysia

Abstract: Problem statement: Handwriting identification is the study for identifying or verifying the writer of a given handwritten document. Since the handwriting features are the cornerstone in the writers' classification process, the classifier accuracy is sensitive in terms of how the writers are scored based on the used features. **Approach:** In this study, we introduced swarm intelligence as a features weighting mechanism to differentiate between the features having high importance and those having low importance in the identification process. The weights obtained from the swarm experiments were used to adjust the features scores and then to identify the most important subset feature for the writers classification process. **Results:** The experiments results showed that a significance influence of the feature weights in the handwriting identification process. **Conclusion:** This communication investigated the influence of the feature importance in the handwriting identification process. Binary Particle Swarm Optimization (BPSO) is used as feature selection method and Euclidian Distance (ED) is used as an evaluation function for the BPSO. The BPSO is trained using 956 words of the off-line IAM data (English handwriting) to learn the feature weights. Each word is represented by 29 statistical features.

Key words: Fitness function, feature selection, feature weight, handwriting identification, particle swarm optimization

INTRODUCTION

The ease of handwriting made it the oldest communication medium. This common habit has two main researches handwriting recognition and handwriting identification. Handwriting recognition is the task of determining the meaning of a handwritten text by transforming a language represented in graphical marks into a symbolic representation. Handwriting identification is the study for identifying or verifying the writer of a given handwritten document. Handwriting identification is a relatively new area of handwriting research when compared to the handwriting recognition or signature verification areas. Handwriting features are characteristics useful for writer discrimination (Srihari *et al*., 2005). Analysis of allographs (characters) and allograph combinations (words) is the key for obtaining those discriminating feature (Zhang and Srihari, 2003). Several studies have shown that handwritten elements are not equal in their discriminating power.

Handwritten words carry more individuality than handwritten allographs (Zhang and Srihari, 2003; Tomai *et al*., 2004). Capital letters bear more individual

information than simple characters like "i" or "c" is proven by (Pervouchine and Leedham, 2006). Principal Component Analysis (PCA) is used by (Wang and Ding, 2004) for reduced the dimensionality of the feature space based Chinese handwriting. Aimed to improve the identification accuracy (Schlapbach *et al*., 2005) has evaluated the set search (Kittler, 1978) feature selection algorithms and GA on the problem of writer identification. The study has concluded that feature selection can significantly improve the writer identification rate using a substantial smaller set of features. Writer relevant features based artificial immune systems is presented in (Muda and Shamsuddin, 2005). Chapran (2006) has described a writer identification system that can be used in a resource contained embedded environment. Since the embedded environment has to have minimum computational costs and minimum identification errors, a feature selection algorithm based on likeness coefficients is proposed. Genetic Algorithm (GA) is used by (Gazzah and Amara, 2006) to select subset features for writer identification using Arabic script. Particle Swarm Optimization (PSO) has been used by (Das and Dulger, 2007) to select a reliable signature

Corresponding Author: Khaled Mohammed Bin Abdl, Faculty of Applied Science,
Hadramout University for Science and Technology, Yemen

verification rate (fine tuning between acceptance and rejection rates). Pervouchine (2006) has shown that, not every document examiner feature can easily be represented as a computational feature and vice versa. Writer invariant features have been studied by (Bensefia *et al*., 2002; Muda *et al*., 2008).

Since, the handwriting features are the cornerstone in the identification process, the classifier accuracy is sensitive in terms of how the writers are scored based on these features. This communication is intended to investigate the effect of the feature selection and the feature weight in the handwriting identification problem. The PSO (Kennedy and Eberhart, 1995) has ability to perform such role and learn the feature weights. Since, PSO works on local level (particle level) and global level (swarm level), where many solutions are suggested for the problem and the best solution among them is selected. Furthermore, PSO is still not tested for handwriting identification, but it could revealed a high performance in some related fields like pattern classification (Tu *et al*., 2006; Huang and Kechadi, 2006), signature verification (Das and Dulger, 2007), handwriting digit recognition (Sahel Ba-Karait and Shamsuddin, 2008).

MATERIALS AND METHODS

Particle Swarm Optimization (PSO): Intelligence System (SI) is the collective intelligence resulting in the collective behaviors of (unsophisticated) individuals interacting locally and with their environment causing coherent functional global patterns to emerge (Ahmed, 2004). Particle Swarm Optimization (PSO) which is inspired by the social behavior of bird flocking or fish schooling and Ant Colony Optimization (ACO) which is inspired by behavior of ants are the primary computational parts of swarm intelligence.

Particle Swarm Optimization (PSO) is a population-based stochastic search algorithm. The PSO like other evolutionary algorithms (e.g., genetic algorithm) performs searches using a population (called swarm) of individuals (called particle) that are updated from iteration to iteration. Compared to GA, PSO is fast to implement since it has no evolution operators such as crossover and mutation i.e., few parameters to be adjusted.

The PSO was originally developed by (Kennedy and Eberhart, 1995); it made up of three parts a momentum, a cognitive, and a social. The momentum part consists of the V_{id} term to keep the particle moving without getting trapped. The cognitive part is P_{id}, each

particle keeps track of its best location which is associated with the best solution the particle has achieved so far. The best position ever encountered by all particles of the swarm is also communicated to all particles. The social P_{gd} term represents the information shared between all particles.

Generally, the initially start of PSO is created by distribute the particles over the search space randomly. Each particle flies through the search space at a velocity that is dynamically adjusted according to two factors its own experience (pBest), as well as according to the experience of all other particles (gBest). In the search space each particle representing a potential problem solution. At any iteration, each particle updates its own velocity and its position using Eq. 1 and 2 respectively. Finally, after several iterations the optimized (optimal or near optimal) solution is found. Figure 1 summarizes the work mechanism of PSO-ED:

$$V_{id}(t+1) \leftarrow w*V_{id}(t)+c_1 r_1 (P_{id}(t)-X_{id}(t))+c_2 r_2 (P_{gd}(t)-X_{id}(t)) \quad (1)$$

Where:

$V_{id}(t)$	=	The velocity of the particle i in the time point t in the search space along the dimension d
$p_{id}(t)$	=	The best position in which the particle previously got high fitness value, it is called pBest.
$x_{id}(t)$	=	The current position of the particle i in the search space
r_1 and r_2	=	Random generated numbers in the range (0,1)
$p_{gd}(t)$	=	The overall best position in which a particle got best fitness value, it is called the gBest
c_1 and c_2	=	Acceleration parameters
W	=	Inertia weight, its value is decreased linearly over the time from 0.9-0.4:

$$X_{id}(t+1) \leftarrow X_{id}(t)+ V_{id}(t+1) \quad (2)$$

Where:

$x_{id}(t+1)$	=	The new position which the particle I must move to, where
$x_{id}(t)$	=	The current position of the particle i
$V_{id}(t+1)$	=	The new velocity of the particle I resulting in the calculation in Eq. 1 which mainly determines the new position of the particle

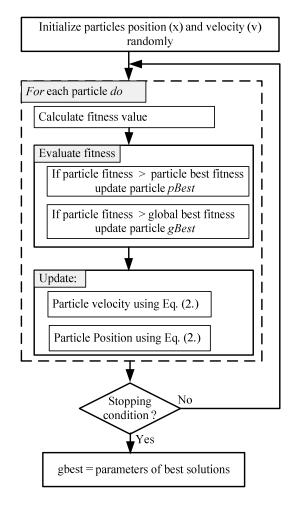The velocity of the particle must be in the range $[V_{max}, V_{min}]$.

Fig. 1: Flowchart the PSO-ED work mechanism

Initially, PSO was set up as an optimization technique for real-number spaces. The binary PSO (Kennedy and Eberhart, 1997) is extension of the continuous PSO, since many optimization problems occur in a space featuring discrete, qualitative distinctions between variables and between levels of variables. in which the particle position is represented as bit string rather than real numbers. In BPSO the velocity is became a probability based a sigmoid function Eq. 3 that transform velocity values randomly into range (0 or 1):

$$x_{ij}(t+1) = \begin{cases} 0 \text{ if } p_{ij}(t) \geq \dfrac{1}{1+e^{-vij(t)}} \\ 1 \text{ otherwise} \end{cases} \quad (3)$$

Handwriting features: The features that have been used in this paper are 29 statistical (moment and words'

measurements) feature. Invariant moment features displayed by Eq. 4-16, while the words' measurements features which are word area, length, height, upper zone height, middle zone, lower zone height and their relationship e.g., aspect ratio of word length to its width displayed by Eq. 17-22.

Moment features: Moments have been used extensively in computer vision, pattern and handwriting recognition and writer identification. Geometrical moments proved to be most useful to aspects of the shape of handwriting. It was determined that features corresponding to the human perception of word shape can be extracted from two- and three-dimensional moments (Liu *et al.*, 1995). The geometric moment of $(p+q)^{th}$ order of digital image of size M×N is computed by Eq. 4:

$$m_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} x^p y^q f(x,y) \quad (4)$$

where, p, q = 0, 1, 2, ,n; x, y image coordinates and m_{pq} is a geometrical moment of $(p+q)^{th}$ order.

The central moments are computed by Eq. 5:

$$\mu_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} \left(x - \bar{x}\right)^p \left(y - \bar{y}\right)^q f(x,y) \quad (5)$$

where, p, q = 0, 1, 2, ,n; $\bar{x}, \bar{y}$ are x and y means and μ_{pq} is a central moment of $(p + q)^{th}$ order.

Following features are extracted from second and third order moments:

$$\theta = \tan^{-1} \frac{\mu_{02} - \mu_{20} + \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}}{2\mu_{11}} \quad (6)$$

$$\lambda_1 = \frac{(\mu_{20} + \mu_{02}) + \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}}{2} \quad (7)$$

$$\lambda_2 = \frac{(\mu_{20} + \mu_{02}) - \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}}{2} \quad (8)$$

where, p, q = 0, 1, 2, ,n; λ_1, λ_2 are the inertial moments, and μ_{pq} is a central moment of $(p + q)^{th}$ order.

The orientation feature is computed by Eq. 9:
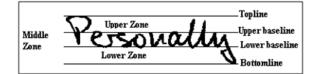
$$f_1 = \frac{\theta}{\pi/2} \quad (9)$$

Fig. 2: English writing zones

The inertial ratio feature is computed by Eq. 10:

$$f_2 = \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2} \qquad (10)$$

The aspect ratio feature is computed by Eq. 11:

$$f_3 = \frac{1}{2}(\frac{\mu_{20} - \mu_{02}}{\mu_{20} + \mu_{02}} + 1 \qquad (11)$$

The spreadness feature is computed by Eq. 12:

$$f_4 = \frac{2\sqrt{(\mu_{20} + \mu_{02})/m_{00}}}{\sqrt{(x_{max} - x_{min})(y_{max} - y_{min})}} \qquad (12)$$

The horizontal skewness feature is computed by Eq. 13:

$$f_5 = \frac{1}{2}(\frac{\mu_{30}^+ - \mu_{30}^-}{\mu_{30}^+ + \mu_{30}^-} + 1) \qquad (13)$$

The vertical skewness feature is computed by Eq. 14:

$$f_6 = \frac{1}{2}(\frac{\mu_{03}^+ - \mu_{03}^-}{\mu_{03}^+ + \mu_{03}^-} + 1) \qquad (14)$$

The balance of horizontal extension feature is computed by Eq. 15:

$$f_7 = \frac{1}{2}(\frac{\mu_{21}^+ - \mu_{21}^-}{\mu_{21}^+ + \mu_{21}^-} + 1) \qquad (15)$$

The balance of vertical extension feature is computed by Eq. 16:

$$f_8 = \frac{1}{2}(\frac{\mu_{12}^+ - \mu_{12}^-}{\mu_{12}^+ + \mu_{12}^-} + 1) \qquad (16)$$

Physically, the $f_1,...,f_8$ features are used to measuring the variance and skewness of black pixels at x, and y coordinates. They represent the extension of an object in the horizontal and the vertical directions.

Word's measurements features: This study extracted 21 statistical feature based a word, such features as area, length, height, height of upper zone, height of middle zone, height of lower zone and their relation e.g. aspect ratio of word length to its height. The following subsections describe how we computed these features.

Area: The area of the word handwriting image was found by summing any black pixels in the image Eq. 17:

$$f_9 = \sum_{x=1}^{M} \sum_{y=1}^{N} f(x,y) \qquad (17)$$

where, f(x,y) represents the black pixel at x, and y coordinates. M and N are the t image dimensions.

Length: The length of word is determined by finding the column number of the first and the last pixel at x coordinate. Then, we subtract the column number of the first pixel from the column number of the last pixel and the result given as the word length:

$$f_{10} = x_{max} - x_{min} \qquad (18)$$

Where:
x_{min} = Represent the column number where the first pixel located, while
x_{max} = The column number where the last pixel located.

Height: The way to find the height of the word image is the same as finding the length of the word image but we looking for row number. Then the height is found by subtracting the row number of the first pixel from the row number of the last pixel in the word image:

$$f_{11} = y_{max} - y_{min} \qquad (19)$$

y_{min} and y_{max} are rows numbers at *y* coordinate.

Upper zone, middle zone, lower zone: As depict by Fig. 2, English writing has three zones upper zone, middle zone and lower zone. To find the heights of each of these three zones. We first, should find the middle zone position in the word image. Then, Eq. 20-22 will be used to calculate the zones heights. In this study the position of the middle zone is determined using the horizontal projection:

$$UpperZoneHeight = Upperbaseline - Topline \qquad (20)$$

$$MiddleZoneHeight = Lowerbaseline - Upperbaseline \qquad (21)$$

$$LowerZoneHeight = Bottomline - lowerbaseline \qquad (22)$$

B1	B2	....	Bn-1	Bn

....

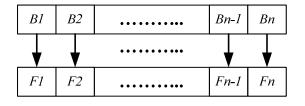| F1	F2	....	Fn-1	Fn

Fig. 3: BPSO particle position representation

where, top and bottom lines represent the first and the last row in the word image respectively. Upper and lower baselines are the middle zone boundaries.

After computing all the features, a normalizing is implemented on the features vector because features can have different scales since they refer to comparable objects.

Particle Position Representation: The BPSO is obtained by changing the position update formula Eq. 2 and 3 while leaving the velocity update formula unchanged. The velocity is became a probability where the value of each bit is retrieved from Eq. 3.

In our case we use the binary PSO in which the particle position is coded to a binary bit string, Fig. 3. This means, each bit can take only the value one or zero which represents the selection case of one feature. When the bit value is 1, its corresponding feature is selected whereas the bit value is 0, it represents the corresponding feature is non-selected.

Evaluation function: Generally, a feature selection model (e.g., PSO) consists of a search mechanism and a fitness function. The search mechanism is used to find subset feature according to a selection criterion, while the fitness function to score the candidates of subset feature. The selection of fitness function depends on the application. In handwriting identification, a classifier is usually selected as a fitness function. In this study we will use the Euclidian Distance (ED) Eq. 23 as an evaluation function for BPSO.

In the BPSO architecture the fitness function is responsible for calculating the value for each particle. The list of particle values at each run consist the weights vector. Each feature weight corresponds to one bit in the particle position, that bit may contain one or zero. The score of the feature is calculated by summing up the feature weights corresponding to the bits containing ones and the feature weights corresponding to the bits containing zeros are excluded. Based on the resulting scores for each feature we can do the handwriting identification in two methods. First, feature selection method, the top n features in the weights ranked list will be selected as an optimal subset feature. Second, feature

weight method, all the features will be used in the identification process. This method differs from the normal process since the features will represent by their importance which corresponding to their weights:

$$ED = \sqrt{\sum_{i}^{n} (r_i - q_i)^2} \qquad (23)$$

Where:
n = Represents the feature number
r_i = Represents the reference document
q_i = Represents questioned document

The Training Procedure: We have set the BPSO variables as follows: number of particles is 5, $V_{max} = 4$, $V_{min} = -4$, $c_1=2$, $c_2=2$, the value of w is in the range [0.9, 0.4], the maximum number of iterations is 500 and the number of runs is 10. In each iteration, each particle selects specific number of features. Based on the selected features, writer identification process is created and evaluated using the fitness function as in Eq. 23. BPSO works as: (1) in the first iteration, the evaluation value of each identification process is selected as pBest for the corresponding particles and the best evaluation value among those five evaluation values is selected as gBest; (2) in the second iteration and above, the new pBest and gBest are selected by comparing the new evaluation values with the previous pBests; (3) by the end of each run, the position of the particle with the gBest value is selected as vector for the best selected features. In step 2 the comparison process is done as: (a) if any new evaluation value is better than the current pBest, the new evaluation value will be selected as pBest; (b) if there is any change in the pBest for any particle, the new pBest will be compared with the current gBest the better one will be selected as new gBest. Finally, the weights of the features are calculated as average of the vectors created in each run. The final feature weights are calculated over the vectors of the feature weights of all writers in the data collection.

RESULTS

The main purpose of conducting this experiment is to investigate the effectiveness of the feature weights in the handwriting identification process. Figure 3 shows the final weights of the features used in this study. Each feature first got weight as average of its selection cases over ten runs for each writer in the data collection (956 English handwriting words). Then the final weight of the feature is average of the feature average weight resulting in the ten runs over the total number of writers.
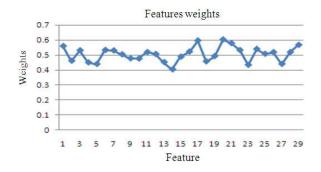
Fig. 4: Feature weights

From the results, we can see that the feature weights are differing in term of the feature importance in the identification process (Fig. 4). For example, quarter of the word's measurement features has got high importance while most of the moment features have got medium importance. Based on the experimental results we can do the handwriting identification in two methods. First, identification based on feature selection, the top n high weight features will be selected as an optimal subset feature. Second, identification based on feature weights, all the extracted features will be used in the identification process after multiplying each feature by its weight. This method differs from the normal process since the features will represent by their importance which corresponding to their weights.

DISCUTION

Among all the phases of handwriting identification, feature selection presents a multi-criterion optimization functions, e.g., simplifying features extracting task, optimizing identification system complexity, reducing running time and improving the classification accuracy. The goal of feature selection is to obtain the significant features by reducing the redundant features. This communication is intended to investigate the effect of the feature selection and the feature weight in the handwriting identification problem. The weights suggested by PSO promoted the scores of the highly important features, which give each handwriting feature the right score it was worth. For our future study, we will apply the feature weights obtained from this study for handwriting identification.

CONCLUSION

In this study, we have investigated the influence of the feature importance in the handwriting identification process. Binary particle swarm optimization is used as feature selection method through training process.

Euclidian Distance (ED) Eq. 23 is used as an evaluation function for the BPSO. We used 956 words of English handwriting for training purpose. The feature weights obtained by the training process have shown a significance influence of the feature weights in the identification process. The final average weights will be used as mechanism to distinguish between the high importance and the low importance features. Future work will be to apply the feature weights obtained from this study for handwriting identification.

REFERENCES

Ahmed, T., 2004. Adaptive Particle Swarm Optimizer for Dynamic Environments. Master Thesis. The University of Texas, Texas.

Bensefia, A., A. Nosary, T. Paquet and L. Heutte, 2002. Writer identification by writer's invariants. Proceeding 8th International Workshop Frontiers in Handwriting Recognition, Aug. 2002, IEEE Xplore Press, USA., pp: 274-279. DOI: 10.1109/IWFHR.2002.1030922

Chapran, J., 2006. Biometric writer identification: Feature analysis and classification. Int. J. Patt. Recog. Arti. Intel., 20: 483-503. http://direct.bl.uk/bld/PlaceOrder.do?UIN=190991830&ETOC=RN&from=searchengine

Das, M.T. and L.C. Dulger, 2007. Off-line signature verification with PSO-NN algorithm. Proceeding of the 22nd International IEEE Symposium on Computer and Information Sciences, Nov. 7-9, IEEE Xplore Press, Ankara, pp: 1-6. DOI: 10.1109/ISCIS.2007.4456842

Gazzah, S. and N.E.B. Amara, 2006. Writer identification using modular MLP classifier and genetic algorithm for optimal features selection. Lecture Notes Comput. Sci., 3972: 271-276. DOI: 10.1007/11760023_40

Huang, B.Q. and M.T. Kechadi, 2006. A fast feature selection model for online handwriting symbol recognition. Proceedings of the 5th International IEEE Conference on Machine Learning and Applications, Dec. 2006, IEEE Xplore Press, Orlando, FL., pp: 251-257. DOI: 10.1109/ICMLA.2006.6

Kennedy, J. and R. Eberhart, 1995. Particle swarm optimization. Proceedings of the IEEE International Conference on Neural Networks. Nov. 27-Dec. 1, IEEE Xplore Press, Perth, WA., Australia, pp: 1942-1948. DOI: 10.1109/ICNN.1995.488968

Kennedy, J. and R.C. Eberhart, 1997. A discrete binary version of the particle swarm algorithm, Proceeding of the IEEE International Conference on System Man and Cybernetics, Computational Cybernetics and Simulation, Oct. 12-15, IEEE Xplore Press, Orlando, FL, USA., pp: 4104-4108. DOI: 10.1109/ICSMC.1997.637339

Kittler, J., 1978. Feature Set Search Algorithms. In: Pattern Recognition and Signal Processing, Chen, C.H. (Ed.). Springer, ISBN: 10: 9028609784, pp: 41-60.

Liu, C.L., R.W. Dai and Y.J. Liu, 1995. Extracting individual features from moments for Chinese writer identification. Proceeding of the 3rd International Conference Document Analysis and Recognition, Aug. 14-15, IEEE Computer Society, Washington DC., USA., pp: 438-441. http://portal.acm.org/citation.cfm?id=844652

Muda, A.K. and S.M. Shamsuddin, 2005. A Framework of Artificial Immune System in Writer Identification. In: BIC`05, Johor Bahru.

Muda, A.K., S.M. Shamsuddin and M. Darus, 2008. Invariants discretization for individuality representation in handwritten authorship. Lecture Notes Comput. Sci., 5158: 218-228. DOI: 10.1007/978-3-540-85303-9_20

Pervouchine, V. and G. Leedham, 2006. Extraction and analysis of document examiner features from vector skeletons of grapheme 'th'. Lecture Notes Comput. Sci., 3872: 196-207. DOI: 10.1007/11669487_18

Sahel Ba-Karait, N.O. and S.M. Shamsuddin, 2008. Handwritten digits using recognition with particle swarm optimization. Proceeding of the 2nd Asia International Conference on Modeling and Simulation, May, 13-15, IEEE Xplore Press, Kuala Lumpur, pp: 615-619. DOI: 10.1109/AMS.2008.141

Schlapbach, A., V. Kilchherr and H. Bunke, 2005. Improving writer identification by means of feature selection and extraction. Proceedings of the 8th International Conference on Document Analysis and Recognition, Aug. 29-Sept. 1, IEEE Xplore Press, USA., pp: 131-135. DOI: 10.1109/ICDAR.2005.139

Srihari, S.N., M.J. Beal, K. Bandi, V. Shah and P. Krishnamurthy, 2005. A statistical model for writer verification. Proceedings of the 8th International Conference on Document Analysis and Recognition, Aug. 29-Sep. 1, IEEE Xplore Press, USA., pp: 1105-1109. DOI: 10.1109/ICDAR.2005.33

Tomai, C.I., B. Zhang and S.N. Srihari, 2004. Discriminatory power of handwritten words for writer recognition. Proceedings of the 17th International Conference on Pattern Recognition, Aug. 23-26, IEEE Xplore Press, USA., pp: 638-641. DOI: 10.1109/ICPR.2004.1334329

Tu, C.J., L.Y. Chuang, J.Y. Chang and C.H. Yang, 2006. Feature selection using PSO-SVM. IAENG Int. J. Comput. Sci., 33: 138-143. http://www.iaeng.org/IJCS/issues_v33/issue_1/IJCS_33_1_18.pdf

Wang, X. and X. Ding, 2004. An effective writer verification algorithm using negative samples. Proceedings of the 9th International IEEE Workshop on Frontiers in Handwriting Recognition, Oct. 26-29, IEEE Xplore Press, USA., pp: 509-513. DOI: 10.1109/IWFHR.2004.14

Zhang, B. and S.N. Srihari, 2003. Analysis of handwriting individuality using word features. Proceedings of the 7th International Conference on Document Analysis and Recognition, Aug. 3-6, IEEE Computer Society, Washington DC., USA., pp: 1142.