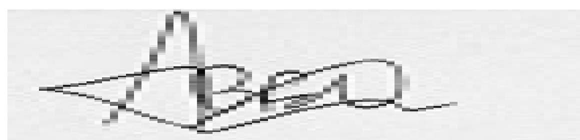
OFFLINE SIGNATURE IDENTIFICATION USING STATISTICAL AND MODEL-BASED APPROACH

Padmajadevi. G[1], Dr. Aprameya K.S[2]
[1]Associate Professor, MCE, Hassan, INDIA
[2]Research Guide, UBDT, Davanageri, INDIA

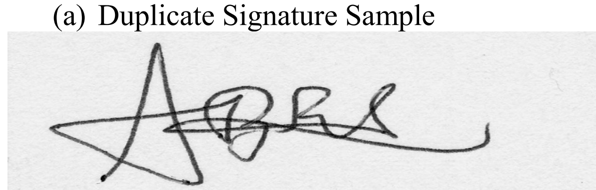
Abstract
The main task of forensic signature analysis is the identification of the writer of handwritten signature. An approach to machine intelligence is based on statistical modelling of data. With a statistical model in hand, one applies probability theory and decision theory to get an algorithm. The statistical model one uses is crucially dependent on the choice of features. Hence it is useful to consider alternative representations of the same measurements. For the statistical approach we evaluate multi-scale statistical features and find one of them to improve the identification performance of single scale features. Model based approach involves s the use of pre - defined models of small strokes of handwriting called graphemes.
Keywords: signature identification, graphemes, edge-hinge, principal component analysis (PCA), clustering.

1.Introduction

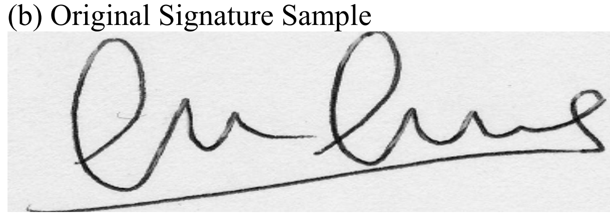
The main task of forensic handwritten signature analysis is the identification of the writer of handwritten signature. An example of handwritten signature by two different writers is shown in Figure 1. Usually, the signature has to be assigned to one of a list of writers, e .g. suspects in a criminal case. Currently, writer / signature identification is performed by forensic handwriting experts. A recent study revealed that the judgements of these experts lack reliability [2]. The important, sometimes even decisive, role that these judgements play in criminal courts prompts for a more objective way of handwriting analysis. Artificial intelligence offers approaches for realising the automatic assignment of handwritten signature to a writer. We distinguish two main approaches to automatic handwritten signature analysis: (1) the statistical approach and (2) the model- based approach. The statistic al approach entails a statistical analysis of features extracted from the handwritten signature [1]. The model- based approach involve s the use of pre -defined models of small stroke s of handwriting signature called graphemes [6].



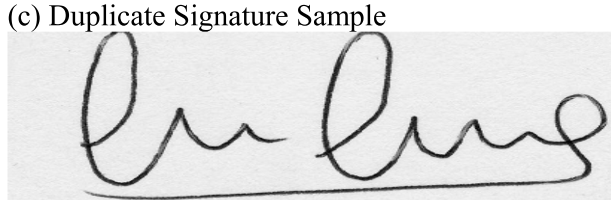
(a) Duplicate Signature Sample



(b) Original Signature Sample



(c) Duplicate Signature Sample



(d) Original Signature Sample

Fig 1: Sample signatures of MCYT Database

2. Methodologies to writer identification

In this area we talk about the statistical approach and the model-based way to deal with essayist recognizable proof. Both the statistical and the model-based methodologies comprise of two phases: a feature extraction stage and classification stage. In the feature extraction stage, features are extricated from handwriting and are stored in feature vectors. In the classification stage, the feature vectors are mapped onto classes representing the writers.

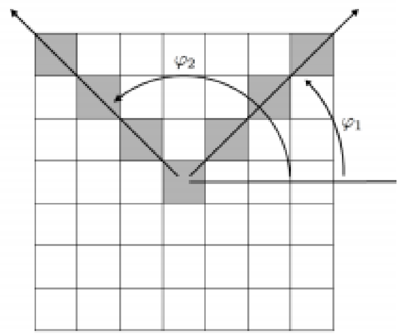


Fig 2: Angle pair $p(\phi 1 , \phi 2)$

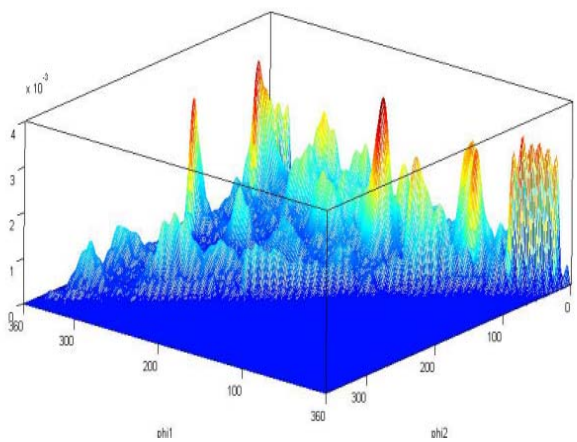


Fig 3: Three-dimensional histogram plot of the edge-hinge distribution showing the frequency of occurrence of angle pairs.

2.1 Statistical approach

Research in automatic writer identification of handwritten signature has mainly focused on the statistical approach. This has prompted the particular and extraction of statistical elements, for example, run-length distributions, inclines distribution, entropy, and edge-pivot distribution. An overview of statistical features is given by Bulacu *et al.* [1]. They found that the edge-hinge distribution feature outperforms all other statistical features. Therefore, we focus on this feature.

Edge-hinge distribution is a feature that characterizes the changes in direction of a writing stroke in handwritten signatures. The edge-hinge distribution is extracted by means of a window that is slid over an edge-detected binary handwriting image. Whenever the central pixel of the window is *on*, the two edge fragments (i.e. connected sequences of pixels) emerging from this central pixel are considered. Their directions are measured and stored as pairs. A joint probability distribution $p(\phi 1 , \phi 2)$ is obtained from a large sample of such pairs. An example of an angle pair is shown in Figure 2. Figure 3 shows an example of an edge-hinge distribution.

2.2 Model-based approach

The model-based approach relies on a codebook of models of graphemes. Graphemes [6] are small strokes of handwriting signature, which are extracted by applying a robust segmentation algorithm on a handwriting image. Graphemes differ from the edge fragments used for the construction of edge hinge distributions because of the used segmentation algorithm.

3. Refining methodologies to signature identification

In the previous section we have described two approaches to writer identification and identified their limitations. In this section, we propose extensions to both approaches in an attempt to overcome these limitations.

3.1 Improving the statistical approach

For the statistical approach we distinguished one impediment: statistical features are acquired utilizing a single scale. We attempt to overcome this restriction by characterizing two new multi-scale features. The first feature is a variation on the edge-hinge distribution feature. The edge-pivot scattering, as described in section 2.1, is an element measured on a solitary scale. However, characterizations on multiple scales often outperform single-scale characterizations. This thought of multiresolution gives its quality to i.e. wavelet analysis. Therefore, we experimented with combinations of edge- hinge distributions created using different fragment lengths. Second, we performed experiments using wavelet feature s. Wavelet feature s have shown to produce promising results in various digital imaging applications [7]. Additionally, wavelet analysis is a multi-scale analysis technique. However, wavelet analysis has not yet been applied to automatic writer identification. In order to extract wavelet features, first line segmentation is performed using horizontal run- lengths. A 50 × 50 pixel window is slid over the middle of these lines, and wavelet transforms are applied on the contents of this window. A number of wave le t feature s (i.e. coefficients obtained by performing wave le t transforms of different level, type or direction) is selected. On the resulting feature vectors, PCA is performed, conserving approximately 90% of the variance in the data. This results in a 50-dimensional feature vector.

3.2 Improving the model-based approach

In subsection 2.2, we observed that grapheme features are effective writer-specific features. However, their main limitation is that the cons truction of grapheme code books is a time - consuming process. We try to overcome this limitation with grapheme constructed by random selection. Graphemes are extracted in the same way as described by Schomaker *et al.* [6].Instead of training we randomly selected a signature samples from the large set of MCYT database where we stored a samples signatures of different persons. The selected graphemes form the database that is used for the construction of grapheme features.

4. Experiments carried out

We implemented experiments to evaluate the performance of both improved approaches on the *Huge* dataset. This dataset contains of 75 folders where in which each folder contains of 25 samples out of which 15 original samples and 10 fraud samples of various writers and also mixed data samples of 500 numbers.

4.1 Methodology

Our experiments were performed using the following validation method. We trained our classifiers on the folders of each writer's signature samples was collected. The identification performances are estimated using a test set consisting of all datasets which was untrained. The performance of an approach is estimated by its identification performance, which is the percentage of writers that an approach correctly identifies. A higher identification performance means that an approach performs better. Identification performances are measured for various list sizes, where the list size is the number of writers the classifier is allowed to select. In all experiments, classification is performed using a 1-nearest neighbour classifier using chi-square distance. For experiments in the statistical approach we used 1125 distracting writers. In the model-based approach, we used 750 distracting writers.

4.2 Combining statistical and model approach

We combined edge- hinge combinations with features from the model based approach. Combining features, which are produced with the edge hinge combination, increases the identification performance to 97%. Therefore, a larger dataset is necessary to establish a significant improvement of our combined approach over that of Schomaker *et al.* [6].

5. Conversation of this work

Step 1: we observed that introducing multi-scale analysis to the edge-hinge distribution, yielding edge-hinge combinations, improves results with a maximum. This improvement is achieved despite the high dimensionality of the resulting feature vectors. This improvement is caused by the multi-scale property of edge-hinge combinations.

Step 2: We observed that random sample datasets achieve the same results as trained – data samples stored. The time-consuming, self-organizing map training is therefore unnecessary.

Step 3: Last, the wavelet features we examined showed disappointing results. We can probably improve identification performances by using different transforms, such as the Gabor wavelet transform or the banana wavelet transform [4]. However, wavelet features will not be able to equal identification performances as achieved by edge-hinge features or grapheme features.

6. ANN Classifier

ANN Classification is the process of learning to separate samples into different classes by finding common features between samples of known classes. **ANN Classification** is an example of **Supervised Learning**. Known class labels help indicate whether the system is performing correctly or not. This information can be used to indicate a desired response, validate the accuracy of the system, or be used to help the system learn to behave correctly. The known class labels can be thought of as *supervising* the learning process; the term is not meant to imply that you have some sort of interventionist role.

Clustering is an example of **Unsupervised Learning** where the class labels are not presented to the system that is trying to discover the natural classes in a dataset. Clustering often fails to find known classes because the distinction between the classes can be obscured by the large number of features (genes) which are uncorrelated with the classes. A step in ANN classification involves identifying genes which are intimately connected to the known classes. This is called *feature selection* or *feature extraction*. Feature selection and ANN classification together have a use even when prediction of unknown samples is not necessary: They can be used to identify key genes which are involved in whatever processes distinguish the classes.

Table I: The training performance of three network structures

Structure	Number of epochs	Relative Time
Linear Network	1350	7.2%
MLP with 1 hidden layer	1000	65%
MLP with 2 hidden layer	1200	100%

Table II: The classification accuracy of the three implemented neural network training algorithms.

Training Algo	FAR	FRR	OER
Back Propagation	2.5%	2.0%	4.5%
Gradient Descent	1.5%	2.5%	4.3%
Levenberg	2.4%	1.5%	5.0%

***Back-propagation*: ** This is the most widely used algorithm in NN training due to its efficiency, simplicity and performance. Back-propagation has been used successfully in many different environments, and results in the highest classification accuracy compared to the other implemented algorithms. The performance details using back-propagation are presented in Table III (error rates)

Conjugate Gradient Descent: This is the major alternative to back-propagation but is not as widely used in HSV. The conjugate gradient descent algorithm converged much faster during training than backpropagation. Unfortunately, due to the fact that this training algorithm is suited to more complex networks (with several hundred weights), the resulting error rate was somewhat worse than networks trained via back-propagation.

7. Conclusion

As a result, we were unable to show that our improved approaches lead to a better performance. Nevertheless, our research leads to three important new observations.

First, in the construction of a dataset training of a self-organizing map as reported elsewhere is not necessary to obtain high identification performances.

Second, identification performances of the edge-hinge distribution can be improved by combining

edge-hinge features from different window sizes. Third, wavelet features perform worse in the handwriting domain than in other digital imaging domains.

Levenberg-Marquardt: The classification accuracy of the Levenberg-Marquardt training algorithm was slightly better than that obtained using conjugate gradient descent but not as good as back-propagation. However, the training process was exceedingly slow, which may be a problem in a large, dynamic database.

Table III: A summary of signature database used in experiment

Property	Value
Number of genuine signers	1125
Number of forgeries	750
Sample rate	205pps
Resolution	1,000(PPCM)
Error	+/- 0.0127 cm (0.005 inches)

8. Practical Results

Results are shown in Fig.4 – Fog.10.

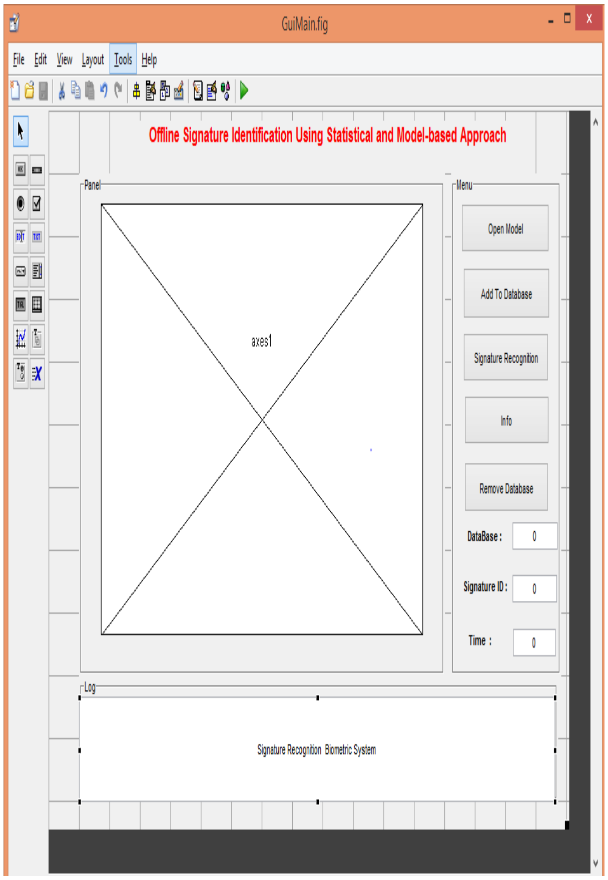


Fig no 4: First slide for experiment

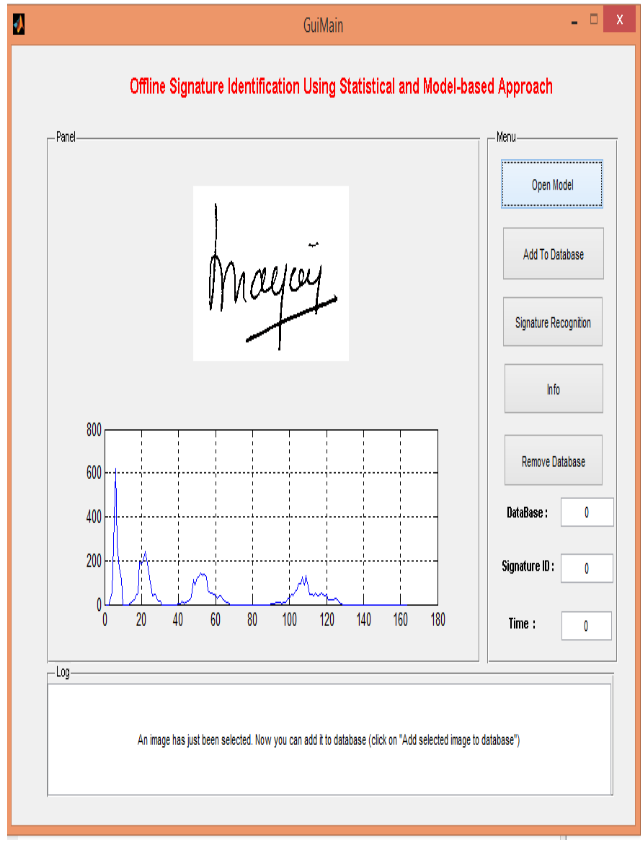


Fig no 6: Add the dataset to database

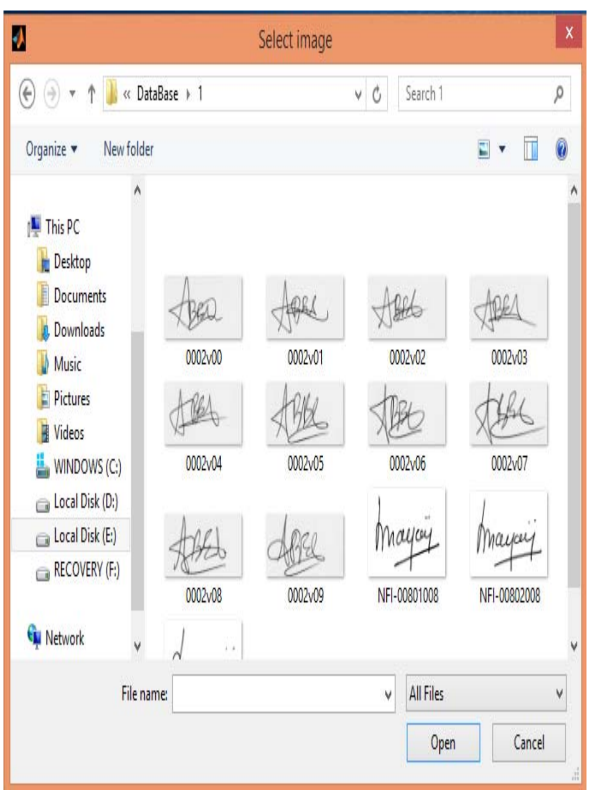


Fig no 5: Select the sample dataset

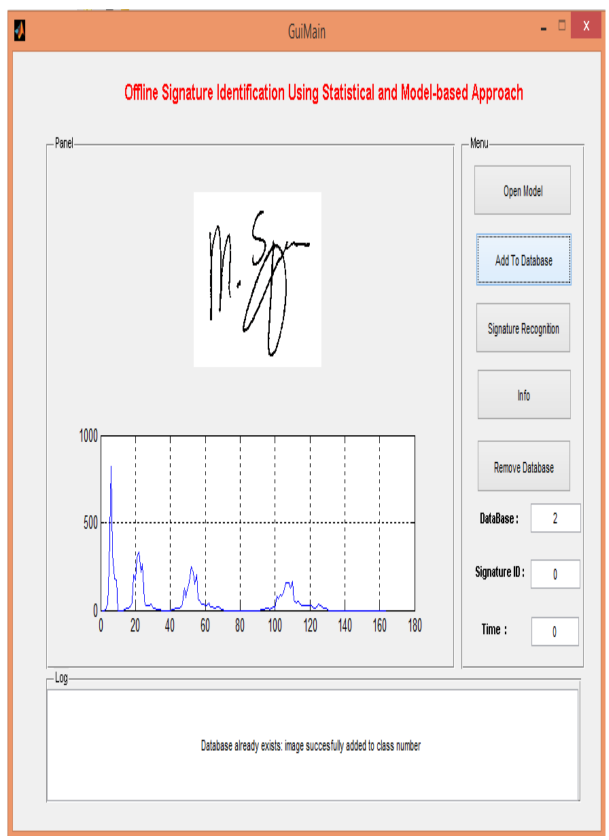


Fig no 7: Sample no 2

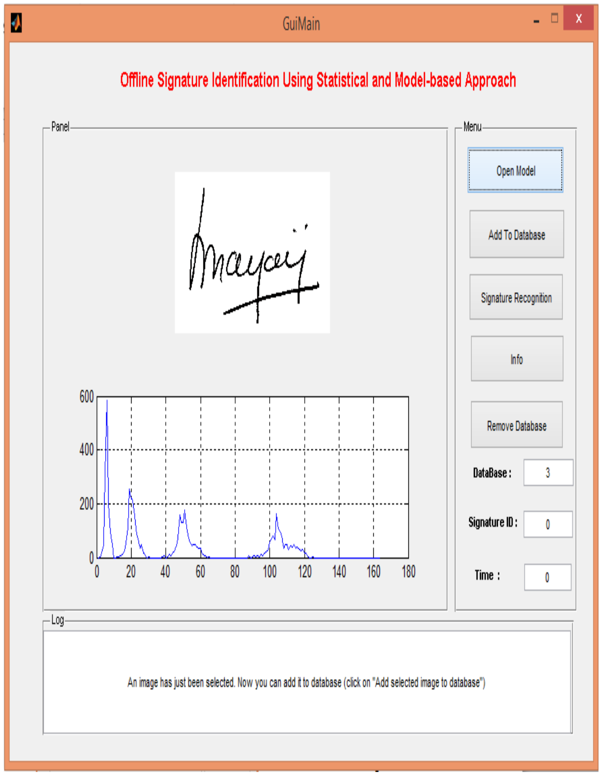


Fig no 8: Select the dataset of sample to recognize

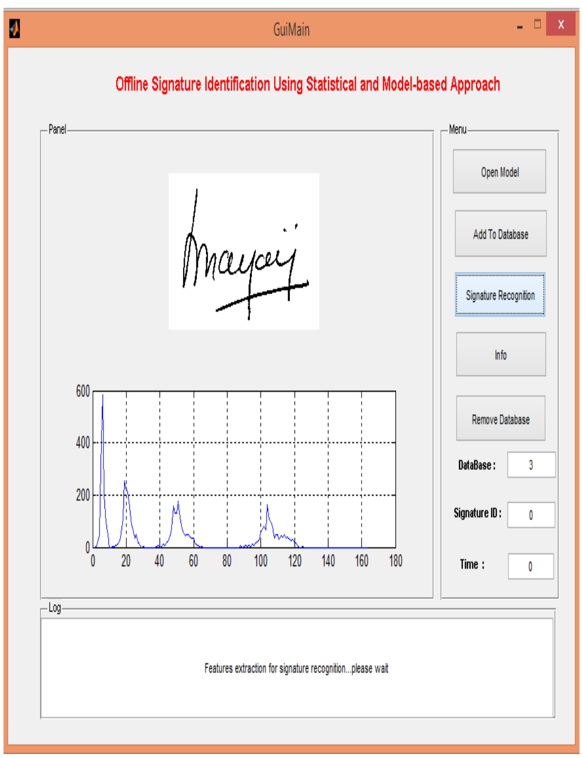


Fig no 9: Processing of signature sample

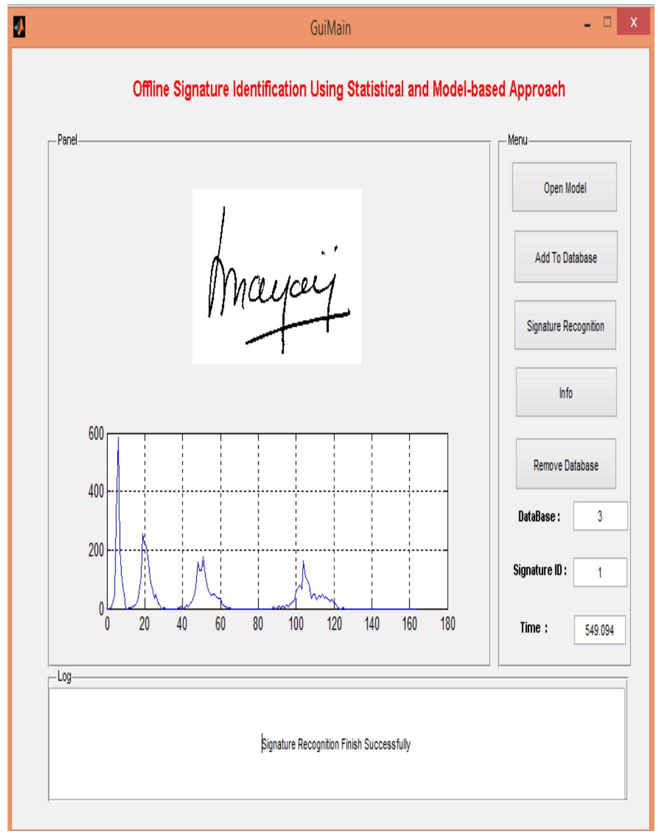


Fig no 10: Exact matching is identified

REFERENCES

[1] M. B ulacu, L. Schom aker, and L. Vuurpijl. Write r identific ation using edge- bas ed direc tional features . In *Proceedings of ICDAR 2003*, pages 937–941, Edinburgh, UK, 2003.

[2] M. Kam, G. Fielding, and R. Conn. Writer identification by professional document examiners. *Journal of Forensic Sciences*, 42:778–785,1997.

[3] T. Kohonen. *Self-organization and associative memory:* 3 *r d edition*. Springer-Verlag New York,Inc., New York, NY, USA, 1989.

[4] G. Peters, N. Kr¨uger, and C. von der Malsburg. Learning object representations by clustering banana wavelet responses. In *Proceedings of the* 1 *st STIPR*, pages 113–118, Prague1997.

[5] A. Schlapbach and H. Bunke. Off-line handwriting identification using hmm based recognizers. In *Proceedings of the* 17 *th ICPR*, pages 654–658, Cambridge, UK, 2004.
[6] L. Schomaker, M. Bulacu, and K. Franke.

Automatic writer identification using fragmented
connected-component contours. In *Proceedings of the* 9 *th IWFHR*, pages 185–190, Tokyo, Japan,
2004.
[7] J.S. Walker. *A Primer on Wavelets for Scientists and Engineers*. CRC Press, Boca Raton, FL, USA, 1999.

[8] . Fang, B., Leung, C.H., Tang, Y.Y., Tse, K.W., Kwok, P.C.K., Wong, Y.K.: Offline signature verification by the tracking of featute and stroke positions. Pattern Recog. 36(1), 91101 (2003)

[9] Fang, B.,Wang, Y.Y., Leung, C.H., TSE, K.W., Tang, Y.Y., Kwok, P.C.K.,Wong, Y.K.: Offline signature verification by the analysis of cursive strokes. Int. J. Pattern Recognit. Artif. Intell. 15(4), 659673 (2001)

[10] Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated
cartography. Communication ACM 24(6), 381395 (1981)

[11] Gilperez, A., Alonso-Fernandez, F. Pecharroman, S., Fierrez, J., OrtegaGarcia, J.: Off-line signature verification using contour features. In: Proceedings of the 11th International Conference on Frontiers in Handwriting
Recognition (2008)

[12] Gonzalez-Rodriguez, J., Fierrez-Aguilar, J., Ramos-Castro, D., OrtegaGarcia, J.: Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems. Forensic Sci. Int. 155(2), 126140(2005)

[13] Impedovo, D., Pirlo, G.: Automatic signature verification: The state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 38(5), 609635 (2008)

[14] Jena, D., Majhi, B., Panigrahy, S.K., Jena, S.K.: Improved offline signature verification scheme using feature point extraction method. In: Proceedings of the 7th IEEE International Conference on Cognitive Informatics, pp. 475480 (2008)

[15] Jonker, R., Volgenant, A.: A shortest augmenting path algorithm for dense and sparse linear assignment problems. Computing 38(4), 325340 (1987).

[16] Liwicki, M., Malik, M.I., Elisa van den Heuvel, C., Chen, X., Berger, C., Stoel, R., Blumenstein, M., Found, B.: Signature verification
competition for online and offline skilled forgeries (sigcomp2011). In: Proceedings of International Conference of Document Image Analysis,pp. 14801485 (2011)

[17] Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60(2), 91110 (2004)

[18] Malik, M.I., Liwicki, M., Dengel, A.: Evaluation of local and global features for offline signature verification. In: Proceedings of the 1st
InternationalWorkshop on Automated Forensic Handwriting Analysis, pp. 3135 (2011)

[19] Malik, M.I., Liwicki, M., Dengel, A.: Evaluation of local and global features for offline signature verification. In: Proceedings of the 1st
InternationalWorkshop on Automated Forensic Handwriting Analysis, pp. 3135 (2011)

[20] Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M.: The det curve in assessment of detection task performance. In:
Proceedings of Eurospeech 97, pp. 18951898 (1997)