SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM

AIVP ADVANCES IN IMAGE AND VIDEO PROCESSING

VOLUME 3, ISSUE 1
ISSN: 2054-7412

Predicting the Identity of a Person using Aggregated Features of Handwriting

[1]Revathi S V and [2]Vijaya M S
PSGR Krishnammal College for women, Coimbatore
sv.revathi15@gmail.com; msvijaya@grgsact.com

ABSTRACT

The identification of an individual based on handwriting is a useful biometric modality. The biometric modalities are broadly classified into two types, namely psychological and behavioral characteristics. The physiological characteristics are fingerprint, face, iris, retina, hand geometry and the behavioral characteristics are voice, signature, gait, handwriting. Handwriting recognition plays vital role in forensic analysis, signature verification and network security. The automatic writer identification will be a valuable and relevant tool in forensic analysis and biometric authentication. Hence it is proposed to design and develop a model for automatic recognition of a person based on handwriting using pattern recognition technique.

Keywords: Classification, Feature Extraction, Prediction, Training, Writer Identification.

1 Introduction

The significance and scope of writer identification has become more prominent in these days. Writer identification is used in most of the areas like digital rights management, forensic expert decision making system and biometric authentication. By combining the writer identification with the authentication system it is used to access the confidential site or data where large amounts of documents, notes, forms and meeting minutes are constantly being processed and managed.

In forensic science writer identification is used to authenticate documents such as records, diaries, wills, signatures and also in criminal justice. The copyrights of electronic media are protected in the digital rights administration system. Two broad categories of biometric modalities are physiological biometrics and behavioral biometrics. The physiological biometrics performs person identification based on measuring a physical property of the human body. The behavioral biometrics use individual traits of a person's behavior for identification. Therefore writer identification is the category of behavioral biometrics.

Writer identification can be generally classified into two types as online and offline. In online, the writing behavior is directly captured from the writer but in offline the handwritten text is used for identification in the form of scanned images. Off-line writer identification is extensively considered as more challenging than on-line because it contains more information about the writing style of a person, such as pressure, speed, angle which is not available in the off-line mode.

Writer identification approaches can be categorized into two type's namely text-dependent and text-independent methods. Dependent on the text content, text-dependent methods only matches the same characters and requires the writer to write the same text consequently.

DOI: 10.14738/aivp.31.997
Publication Date: 28th February, 2015
URL: http://dx.doi.org/10.14738/aivp.31.997

The text-independent methods are able to identify writers independent of the text content and it does not require comparison of same characters. If it requires the same writing content, then this method is not apt for many practical situations. Even though it got a wider applicability, text-independent methods do not obtain the same high accuracy as text-dependent methods do. The basic property of handwriting is that there exists writer invariant which makes writer identification possible. The writer's invariants reflecting the writing style or writing individuality of handwriting can be defined as the set of similar patterns.

2 Literature Survey

Based on the study of various literatures available on writer identification, a brief report is presented in this section about the research directions in writer identification.

In the research work [8], the authors developed the writer identification model using character level and word level. Scanned images of handwritten words were segmented into words and it is further classified into characters. 26 features were used for both word and character. RBF, Polynomial and Kernel were used to train the dataset using SVMLight.

In [9] the author developed a writer identification model using the handwritten documents. The noise in the scanned images is removed by locating the text lines and empty spaces and the height of the text are normalized. The features are extracted using the generalized gaussian density wavelet transform. Weighted Euclidean distance (WED) were used for classification.

Authors in [11] implemented writer identification model in which normalization and Emprical mode decomposition technique was used. Intrinsic Mode function was used to extract the features. IMFs of each function differ from one another. It contains significant information of each handwritten document. The feature vector was classified using K-nearest neighbor.

In [19] the authors proposed writer identification and verified using tamil handwritten words. In scanned images the following operations were performed such as segmentation, noise removal, binarization, edge detection and thinning. Global Features and Local Features are used to extract the features. Three supervised learning algorithms such as Naïve bayes classifier, Decision Tree induction and k-Nearest Neighbor were used for learning the classification model.

In this research work [20] the authors proposed a writer identification model using document images. The normalization technique was applied to normalize the skewed word and features are extracted using the multi-channel Gabor filtering and gray scale co-occurrence matrix. Here two classifiers k-nearest neighbor (KNN) and weighted Euclidean distance (WED) were used for classification.

Author in [21] proposed a writer identification model using document images. The noise in the scanned images was removed by locating the empty spaces and text lines. Gabor, GGD and Contourlet GGD are used to extract the features from the preprocessed image and the performance of the model was evaluated using the features.

In the research work [22] the authors developed a writer identification model using the handwritten document. The noise in the images is removed on both the foreground and background of the image.

The edge based directional feature and edge hinge distribution was used to extract the features from the preprocessed image. K-nearest neighbors (KNN) were used for classification.

From the background study, it was observed that the writer identification problem can be modeled as pattern classification task and can be solved using supervised learning techniques. As machine learning technique can automatically learn the model by taking intelligent hints from the training data and predicts the output more accurately, it has been influenced in this work to pool the various features of handwritings for making the training dataset. Various phases of the proposed implementation are described in Section 3.

3 Proposed Model for Writer Identification

This work aims to develop a discriminative model for identifying a person using the handwritten documents. The language considered here is English and the handwritings have been collected from different writers and preprocessing is done using the normalization technique. The features are extracted using Gabor, GLCM, GGD, Contourlet GGD and directional features. Standard classification algorithms have been employed to build the model and to predict the writer's identity.

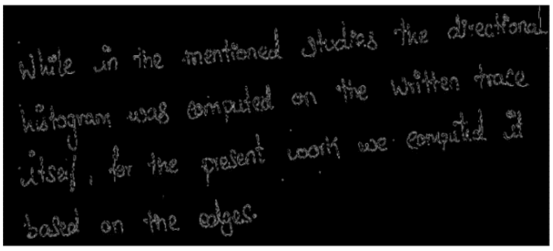
3.1 Data Acquisition

The data acquisition is an important task in writer identification. Text dependent data has been acquired from ten different writers. The paragraphs are scanned using scanner of resolution 300 dpi and a total of 300 JPEG text images are collected from 10 writers at a rate of 30 paragraphs per writer are obtained.

3.2 Preprocessing

Preprocessing is an important task in any mining activity. Here normalization technique is performed prior to preprocessing in order to correct the skewed words in the handwriting image. The space between vertical and horizontal lines has been normalized to produce a well-defined pattern for texture analysis. Then the scanned images are preprocessed to remove the noise and converted into grayscale images to carry further preprocessing tasks. The preprocessing tasks carried out here are edge detection, image dilation and box bounding.

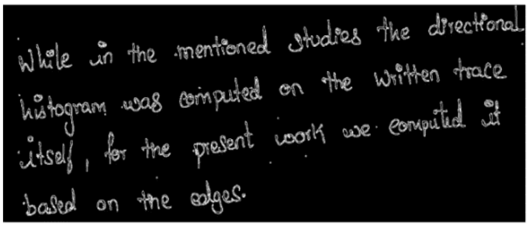
Edge detection: Edges in the binary image are detected using sobel method. The Sobel method finds edges using the Sobel approximation to the derivative. It returns edges at those points where the gradient of I is maximum. Edge ignores all edges that are not stronger than threshold. If threshold do not specified, or if threshold is empty ([]), edge chooses the value automatically.



While in the mentioned studies the directional histogram was computed on the written trace itself, for the present work we computed it based on the edges.

Figure 1: Edge detection

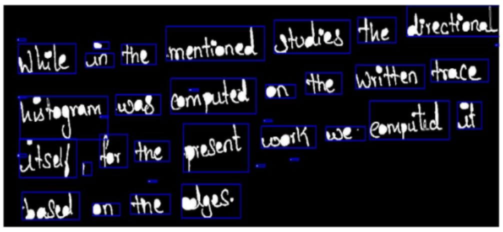
Dilation: The dilation is a fundamental morphological operation. It helps to add the pixels to the boundaries of the images. Based on the size and shape of the structuring element the pixels are added. Once the image is converted into grayscale image, the dilation operation is performed.



While in the mentioned studies the directional histogram was computed on the written trace itself, for the present work we computed it based on the edges.

Figure 2: Dilated image

Box bounding: After the image is converted into dilated image, the pixels in the image are converted into white pixels. To box bound a word, the white pixels are taken for each character with a pixel space of 30 pixels. Once it exceeds the pixel space of 30 then a word is box bounded and it starts from the next word to box bound until it exceeds the space of 30 pixels.



While in the mentioned studies the directional
histogram was computed on the written trace
itself, for the present work we computed it
based on the edges.

Figure 3: Box bounded image

3.3 Feature extraction

Feature extraction plays a vital role in improving the classification effectiveness and computational efficiency. It is used to extract the writer invariance in the form of a pattern. The various features used in preparing the training dataset are described below.

Gabor Filter

The Gabor filter is a bandpass filter used in the image processing for feature extraction. The Gabor filter requires an input image with N*N pixel image along with the frequency (f) and an angle θ. Here the θ and f specifies the location for the Gabor filter. As the size of the image is N*N, the frequency used here is 4, 8, 16 and 32 cycles/degree. The parameters used in this filter are bandwidth, phase shift and lambda.

For each central frequency f, filtering is performed at 0, 45, 90 and 135 degrees. This gives a total of 16 output images for each frequency, from which the writer's features are extracted. These features are the mean and the standard deviation of each output image and the features are extracted from the images.

GLCM

GLCM is a matrix where number of rows and columns is equal to number grey levels G in an image. It is defined over an image to be the distribution of co-occurring values in the given offset. It is a way of extracting second order statistical features. It is used to measure the spatial relationships between pixels. This method is based on the belief that texture information is contained in such relationships. The GLCMs are constructed by mapping the grey level co-occurrence probabilities based on spatial relation of pixels in different angular direction. Greycomatrix function creates the GLCM by calculating how often pixels with grey-level value I occurs horizontally adjacent to a pixels with the value j. Each element (i, j) in GLCM specifies the number of times that the pixels with values I occurred horizontally adjacent to a pixels with value j.

There are 22 texture features associated with GLCM. They are autocorrelation, contrast, correlation, Cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity, maximum probability, sum of squares, sum of average, sum variance, sum entropy, difference variance, difference entropy, information measure of correlation1, information measure of correlation 2, inverse difference homogeneity, inverse difference normalized, Inverse difference moment normalized.

Generalized Gaussian density

The basic idea of the wavelet-based GGD method is to establish corresponding wavelet-based GGD model for a handwriting image. Wavelets are mathematical functions that cut up data into different frequency components, and then study each component with a resolution matched to its scale. The parameters α, β are regarded as the features of the handwritten image. For each handwritten image, the GGD model cut the image using different frequency components are cut into regions. Each region is called as a wavelet subband. For each wavelet subband the probability is maximized for the estimated parameters α, β which are optimal for improving accuracy of writer identification. The probability is estimated as,

$$P\ (\{\alpha,\ \beta\}/X)$$

Each wavelet subband has coefficients as X={x1, x2……}. After the error probability is estimated the likelihood function is defined as,

$$L(X|\{\alpha,\beta\}) = \log\prod_{i=1}^{N} P(x_i|\{\alpha,\beta\}).$$

According to the langrange optimization, the likelihood equation is obtained as follows,

$$\frac{\partial L(x|\{\alpha,\beta\})}{\partial\alpha} = -\frac{N}{\alpha}+\sum_{1}^{N}\frac{\beta|x_i|^{\beta}a^{-\beta}}{\alpha},$$

After the likelihood function is estimated then α, β values are obtained using the following equations,

$$\frac{\partial L(x|\{\alpha,\beta\})}{\partial\beta} = -\frac{N}{\beta}+\frac{N\varphi(1/\beta)}{\beta^2}$$

$$-\sum_{i=1}^{N}\left(\frac{x_i}{\alpha}\right)\log\left(\frac{x_i}{\alpha}\right)$$

where $\varphi(z) = \tau\,(z)/\tau(z)$

$$\hat{\alpha} = \left(\frac{\beta}{N}\sum_{i=1}^{N}|x_i|^{\beta}\right)^{1/\beta}. \qquad \text{--1}$$

The estimation of the β is be calculated by using the following equation,

$$1+\frac{\varphi(1/\hat{\beta})}{\hat{\beta}}-\frac{\sum_{i=1}^{N}|x_i|^{\hat{\beta}}log|x_i|}{\sum_{i=1}^{N}|x_i|^{\hat{\beta}}}$$

$$+\frac{\log((\hat{\beta}/N)\sum_{i=1}^{N}|x_i|^{\hat{\beta}})}{\hat{\beta}} = 0.$$

Once the value of the β is calculated then it is substituted in the equation 1 to find the value of α. By this process α, β values are obtained for each pixel and the features have been obtained for all the input image.

Contourlet Generalized Gaussian density

In GGD, wavelet transform is used to decompose the image into subbands in different frequency and orientation. But wavelet is able to capture only limited directional operation which is an important issue in image analysis. To overcome this problem, multiscale and directional representations can be used to efficiently capture the image's geometrical structures such as edges or contours.

Contourlet transform based on an efficient two-dimensional multiscale and directional filter bank can deal effectively with images having smooth contours. The contourlets are implemented using the double filter bank named pyramidal directional filter bank (PDFB). Laplacian pyramid is used in PDFB to decompose the images into multiscale using 9-7 filter bank. The directional filter bank is used to analyze the multiscale into a four directional subbands. Multiscale and directional decomposition stages in the contourlet transform are independent of each other, because while using PBFB filter it gets a cascade structure. The parameters α, β of the GGD model is taken and used to represent the contourlet subband. The GGD is given as below,

$$p(x;\alpha,\beta) = \frac{\beta}{2\alpha\tau(1/\beta)} exp^{-1(|x|/\alpha)^{\beta}}$$

Where $\tau(.)$ is the Gamma function, i.e.

$$\tau(.) = \int_0^{\propto} exp^{-t} t^{Z-1} dt, Z > 0.$$

Various methods are available to estimate the parameters α and β. Here the maximum likelihood estimation is used to estimate the parameters α, β by converting the subband image into multi-dimensional vector and it is defined as given below,

$$L(x;\alpha,\beta) = log \prod_{i=1}^{L} p(x_i;\alpha,\beta)$$

By using the following equation the features are extracted for the subband images.

$$\frac{\partial L(x;\alpha,\beta)}{\partial\alpha} = -\frac{L}{\alpha} + \sum_{1}^{L} \frac{\beta|x_i|^{\beta}\alpha^{-\beta}}{\alpha}$$

The values of the GGD are taken as input and the filters mentioned above are applied to extract the features from a given input image. In this manner the features are obtained for all the images.

Directional features

The normalized feature vector for classification is obtained using the directional features here. First the input image is characterized into four types, such as vertical line, horizontal line, left diagonal and right diagonal. The values are calculated from the four directions, as fraction of the distance traversed across the image. If the transition is computed from left to right, a transition found close to the left is assigned a high value compared to a transition computed further to the right. A maximum value (MAX) is defined as the largest number of transitions that is recorded in each direction. If there are less than MAX transitions recorded, then the remaining MAX transitions are assigned values of 0.

The transition value is calculated for a particular direction. To calculate the directional transition, the transition value is divided by a predetermined number. Here the predetermined number is 10. Thus eight features are obtained for one transition. Then this process is repeated for the remaining transitions and the features are obtained, using the following formula,

nrFeatures * nrTransitions * nrVectors * resampledMatrixHeight (Width)

The above features are extracted by developing MATLAB code and feature vectors are generated for all the input images. Particle swarm optimization (PSO) method is used to select the contributive feature and to reduce the dimension of the feature vector.

Feature Selection

The feature selection method used here is particle swarm optimization (PSO) method. The PSO method is used to reduce the number of features obtained during the feature extraction. The feature selection is done to in order to speed up the processing rate and predictive accuracy. The features are extracted for each and every pixel for a given input image. So to reduce the dimension of feature vector, feature selection method is used. If there are n number of features, then the threshold value of the feature selection is n<10. The features are reduced based on the weight assigned to each feature. After the weight is assigned to each feature, the features are selected in the descending order based on the weight of the features.

4 Experiments and Results

The writer identification model proposes the application of computational intelligence technique to develop discriminative model for a person identification using their handwriting. The language considered here is English and the handwritings have been collected from 10 persons as described in section 3(A). Scanned images of handwritten documents are used to prepare the dataset. The scanned images are normalized as explained in section 3(B). Features like Gabor, Gray level co-occurrence matrix, generalized gaussian density, Contourlet generalized gaussian density and Directional features are taken into account as described in the section 3(C).

Since the total number of handwritten document images considered is 300, the training dataset with 250 instances is developed. Three independent experiments have been carried out with three different datasets and by training the standard classification algorithms naive bayes, decision tree, KNN, support vector machine and neural networks. The dataset is prepared in .CSV format in order to classify the instances using R. The built classifiers have been evaluated using hold out method and the results are analyzed.

For the first experiment, the features associated with Gabor filter, Gray level co-occurrence matrix, generalized gaussian density and Contourlet generalized gaussian density are aggregated and the training dataset is developed. The dataset is then learned using above mentioned classification techniques and their performance is evaluated using measures such as accuracy, precision, recall and F measure. The results of the experiment are shown Table I and illustrated in Figure 1.4.

Table 1: Results of classifiers based on GGD and Contourlet GGD features

Alg	Accuracy	Precision	Recall	Fmeasure
NB	63%	0.723	0.515	0.638
DT	85%	0.912	0.749	0.852
KNN	82%	0.901	0.721	0.823
NN	69%	0.689	0.678	0.699
SVM	75%	0.735	0.755	0.744

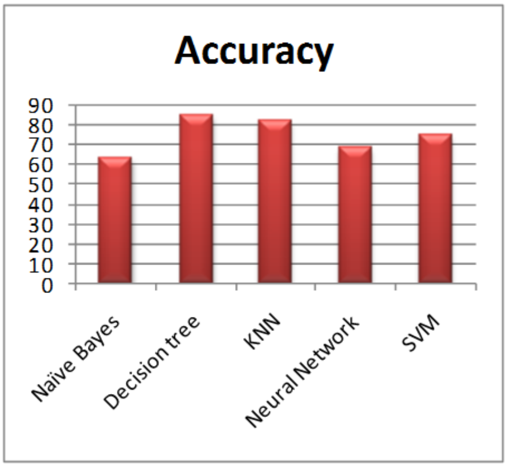


Figure 4: Performance of classifiers based on GGD and Contourlet GGD features

To carry out the second experiment, along with texture features i.e. Gabor and GLCM, directional features are combined to develop training dataset. The dataset is then trained and the classifiers are evaluated. The results of the classifiers in terms of accuracy, precision, recall and F measure are shown in Table II and illustrated in Figure 1.5.

Table 2: Results of classifiers based on Directional features

Alg	Accuracy	Precision	Recall	Fmeasure
NB	70%	0.826	0.635	0.719
DT	61%	0.768	0.524	0.618
KNN	77%	0.723	0.756	0.788
NN	70%	0.711	0.704	0.707
SVM	60%	0.725	0.512	0.663

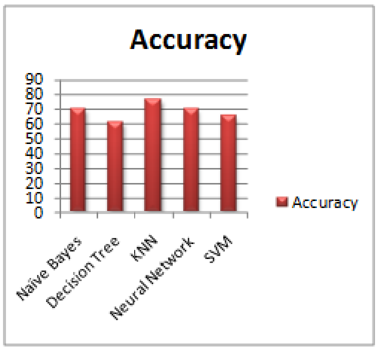


Figure 5: Performance of classifiers based on Directional features

In the third case, all five categories of features are pooled to form a training dataset. The results of learning the classifiers are shown in Table III and illustrated in below.

Table 3: Results of classifiers based on aggregated features

Alg	Accuracy	Precision	Recall	Fmeasure
NB	85%	0.895	0.876	0.852
DT	86%	0.923	0.756	0.861
KNN	92%	0.922	0.927	0.929
NN	80%	0.974	0.715	0.823
SVM	88%	0.865	0.871	0.867

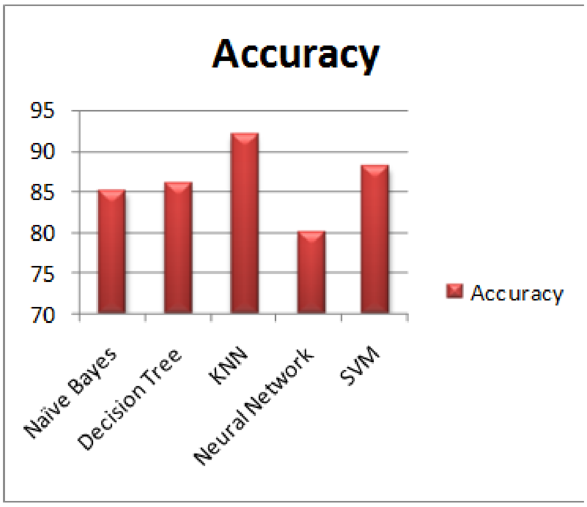


Figure 6: Performance of classifiers based on aggregated features

From the above experiments it was observed that the performance of the classifiers is high when training dataset contains aggregated features. The classification models built using GGD and Contourlet GGD features produced an accuracy of about 82%. The classification models built using directional features produced an accuracy of about 77%. When all the features Gabor, GLCM, GGD, Contourlet GGD and directional features are combined together, the models showed an accuracy of about 92%. Hence it is concluded that the pooled features can better produce a trained model for accurate writer identification.

5 Conclusion

This paper demonstrates the modeling of writer identification as classification task and describes the implementation of supervised learning approach for identifying the writer based on their handwriting. Various features such as Gabor, GLCM, GGD, Contourlet GGD and directional features are pooled to develop the training datasets and are used to fabricate the models. The outcome of the experiments proves that, the writer identification model is effectual when the collective features are used in learning. Further work can be extended for text independent handwriting.

REFERENCES

[1] Eibe Frank, Ian H. Witten. (2005), Data Mining – Practical Machine Learning Tools and Techniques. Elsevier Gupta GK “Introduction to Data Mining with Case Studies”.

[2] Mitchell T. “Machine learning”, Mc Graw-Hill International edition.

[3] Sreeraj.M and Sumam Mary Idicula. (2011), “A Survey on Writer Identification Schemes”, International journal of computer applications, Vol. 26, No. 2.

[4] Marius Bulacu, Lambert Schomaker, Louis Vuurpijl. (2003), “Writer Identification Using Edge-Based Directional Features”, Proceedings of the Seventh International Conference on Document Analysis and Recognition.

[5] Saranya K, Vijaya M.S (2013) “An interactive tool for writer identification based on offline text dependent approach”, International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No. 1.

[6] Plamondon, R., Lorette, G. (1989) Automatic Signature Verification and Writer identification ―The State of the Art, ‖ Pattern Recognition, vol. 22, no. 2, pp. 107-131.

[7] H. E. S. Said1, G. S. Peake1, T. N. Tan2 and K. D. Baker1 “Writer Identification from Non-uniformly Skewed Handwriting Images”.

[8] Saranya K, Vijaya M.S (2013) “Text Dependent Writer Identification using Support Vector Machine”, International Journal of Computer Applications (0975 8887) Volume 65 No.2

[9] Zhenyu, H., Xinge, Y., Tang, Y.Y. (2008) “Writer Identification using global wavelet-based features” neurocomputing 71, 1832–1841.

[10] Plamondon, R., Lorette, G. (1989) Automatic signature verification and Writer identification—The State of the Art, Pattern Recognition, vol. 22, no. 2, pp. 107-131

[11] Karukara K and Dr. B.P. Mallikarjunasamy “Writer Identification based on offline handwritten Document images in Kannada language using Emprical mode decomposition method” Volume 30– No.6, September 2011.

[12] Zhu, Y., Tan, T., Wang, Y. (2001) Font Recognition Based on Global Texture Analysis, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 10, pp.1192- 1200.

[13] Zhang, B., Srihari, S. (2003) Analysis of Handwritten Individuality Using Word Features, Proc. Seventh Int‘l Conf. Document Analysis and Recognition (ICDAR), pp.1142-1146.

[14] Bensefia, A., Paquet, T., Heutte, L. (2005) ―A Writer Identification and Verification system, Pattern recognition system, vol. 26, no. 10, pp. 2080-2092

[15] Bensefia, A., Nosary, A., Paquet, T., Heutte, L. (2002) Writer Identification by Writer‘s Invariants,Proc. Eighth Int‘l Workshop Frontiers in Handwriting Recognition, pp. 274-279.

[16] Marti, U.V., Messerli, R., Bunke, H. (2001) ―Writer Identification Using Text Line Features, Proc. Sixth Int‘l Conf. Document Analysis and Recognition (ICDAR), pp. 101- 105.

[17] Hertel, C., Bunke, H. (2003) ―A Set of Novel Features for Writer Identification, Proc Fourth Int‘l Conf. Audio and Video-Based Biometric Person Authentication, pp. 679-687.

[18] Schlapbach, A., Kilchherr, V., Bunke, H. (2005) Improving Writer Identification by Means of Feature Selection and Extraction, Proc. Eighth Int'l Conf. Document Analysis and Recognition (ICDAR), pp. 131-135.

[19] Thendral, Vijaya.M.S., "Supervised Learning Approach for Tamil Writer Identity Prediction using Global and Local Features".

[20] H. E. S. Said1, G. S. Peake1, T. N.Tan2 and K. D. Baker1 "Writer Identification from Non-uniformly Skewed Handwriting Images".

[21] Z.Y.He, " A Contourlet based method for writer identification"

[22] Marius Bulacu, Lambert Schomaker, Louis Vuurpijl. (2003), "Writer Identification Using Edge-Based Directional Features", Proceedings of the Seventh International Conference on Document Analysis and Recognition