Design of Offline and Online Writer Inference Technique

R. Raja Subramanian, Ramalakshmi Ramar

Abstract**: *Writer inference systems tend to identify and verify the authorship of the handwritten documents. Each writer will have his own style of writing that uniquely identifies the writer. Hence authorship identification finds its application in forensic document analysis. It is also considered as one of the biometric features of a person, so helps in security to uniquely identify a person. Recognition of writers online has its application in detecting the identity thefts. That is compromising one's social media account and sending messages to others as if he were an authentic sender. By discriminating the writing characteristics of the original and intruder, the masquerader can be identified. In this survey various works contributing to feature extraction and prediction of writers are discussed.***

Keywords : Authorship identification, Run length features, Image transforms, Writer prediction.

I. INTRODUCTION

Authorship identification can be of two types namely, online writer identification and offline writer identification. Online mode of writer inference enables identification and verification of writers while the document is being written. It makes use of online features including pen strokes, stroke length, writing speed, curls and angles of pen. On the contrary, offline writer inference leverages pattern recognition algorithms for recognizing authors of the handwritten document in question. Structural and textural features of the handwriting images are used for analysis.

The paper is structured in the following manner: Section 2 deals with feature extraction techniques producing descriptors, uniquely describing writers, are discussed, Section 3 deals with various machine learning techniques for predicting the writer, Section 4 contains the overall methodology of authorship identification followed by Conclusion and scope for future work.

II. FEATURE EXTRACTION

Djeddi (2013) [1] proposed an encoding based algorithm for writer inference. This offline writer inference algorithm encodes the binary handwritten image to create run lengths. The binary handwritten image is subjected to four-directional scans to compute a run length pattern. The run lengths are capable of describing the features of the corresponding handwritten document. The feature extraction process is illustrated in Fig. 1. It is a 8 × 6 pattern with binary values {0,1}. The image represents a typical shape of English character 'M'.

The values in the matrix indicates the count of runs of length 1, 2, 3, 4, 5, 6 in three directions 45^{O}, 135^{O} and 180^{O} depicted in fig 1b, 1c and 1d respectively. The 90^{O} direction accounts run lengths of 7, 8 in addition, as shown in fig. 1e. The first row of the matrix represents the number of zeros, with first element representing the isolated 1, second depicting for two consecutive 0s, and so on. The second row replicates the same configuration for the value 1. The encodings of the runs from these matrices are normalized to construct a histogram, serving as a descriptor for the corresponding handwritten document.

This method is not useful when the handwriting is cursive which is the actual case in real time. The run length matrices cannot provide appropriate descriptive features in such cases. Surinta (2015) [2] used bagging techniques leveraging gradient features including Histogram of Oriented Gradients (HOG) and a popular pattern recognition algorithm, Scale Invariant Feature Transform (SIFT) descriptor to characterize the writers. In HOG, an image is divided into several small grids. The horizontal and vertical component of the gradients of each grid is used to compute the histogram. The histograms of the grids are bagged and normalize. The normalized histogram serves as a potential descriptor of the handwritten document.

1	**0**	**0**	**0**	**0**	**1**
1	**1**	**0**	**0**	**1**	**1**
1	**0**	**1**	**1**	**0**	**1**
1	**0**	**0**	**0**	**0**	**1**
1	**0**	**0**	**0**	**0**	**1**
1	**0**	**0**	**0**	**0**	**1**
1	**0**	**0**	**0**	**0**	**1**
1	**0**	**0**	**0**	**0**	**1**

(a)

Revised Manuscript Received on December 16, 2019.

***** Correspondence Author

R. Raja Subramanian*, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Virudhunagar, India. Email: rajasubramanian.r@klu.ac.in

Ramalakshmi Ramar, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Virudhunagar, India. Email: rama@klu.ac.in

4	3	2	3	0	0
15	1	1	0	0	0

(b)

4	3	2	3	0	0
15	1	1	0	0	0

(c)

2	1	0	6	0	0
14	3	0	0	0	0

(d)

2	2	0	4	2	2	0	0
4	0	0	0	0	0	0	2

(e)

Fig. 1. Encoding of binary runs (a) a binary (0/1) image depicted in 8×6 matrix (b) encodings in diagonal angle 45^o (c) encodings at 90^o (d) encodings at 135^o (e) encodings at 180^o.

SIFT resembles HOG in some aspects, but has some significant differences. SIFT identifies key points in the image. Key points are those that exhibit extreme deviations with small deviations. If (x,y) represents a point in the image, the displacement (Δx,Δy) of the point would create a considerable deviation in pixel intensity. The intensity deviation is characterized based on the threshold λ, fixed by leveraging the sensitivity of the application. These kinds of descriptors are robust against translations, rotations and distortions in the image. The image is smoothened by convoluting it with Gaussian window of dimension 4×4. The histograms are computed for each window in 8 bins covering 360^o. The orientation histogram is strengthened using the gradient weights. These oriented histograms with gradient weights are effectively used as descriptors, of handwritten images, characterizing the writers.

HOG captures only global features and cannot serve as a robust feature descriptor. Though SIFT captures local features, it is quite slow and does not work well with blurs in images. Hence, there is a need to have a local descriptor that uniquely describes the handwritten image and also quite quickly. Sheng He (2015)[3] proposed a method to detect junction points in the image. The junction points are usually corner points, the special case of key points in SIFT. The detected junctions form the codebook based representation for the handwritten image called Junclets. First the candidate junctions are detected then branch strength of the junction, number of branches and junction orientation are computed. The junction is used as a local descriptor of the handwritten image. The junction points do not characterize the shape of the image. Hence it cannot serve as the only descriptor for effective writer identification.

Ooia (2016) [4] used Radon Transform for feature extraction from images. Discrete Radon Transform (DRT) signifies the image projection at various angles. It converts lines in two dimensional images into set of line parameter values. Each line is capable of representing a peak positioned by its line parameter. After this transformation, the transform values were produced as result. The DRT of an image is computed in the following manner: Let the number of pixels in the image be M, and the pixel intensity of the ith pixel is denoted by I_i, where i = 1…N. The DRT vector is calculated leveraging beams per angle that do not overlap each other and on the whole there are ψ angles. The pixels' cumulative density within the jth beam is characterized by Y_j, j = 1…$\alpha\psi$. This is called the beam sum at the beam j. The radon transform is its discrete form is evaluated as in (1).

$$Y_j = \sum_{i=1}^{M} w_{ij}\, I_j \qquad (1)$$

where j =1, 2… $\alpha\Omega$, the weighted combination of the pixel i to the jth beam sum is denoted by w_{ij}. Two dimensional interpolations are used to determine the value of $w_{ij.}$ The beams are projected at various angles along the image. The accuracy of Radon transforms, as stated in (1) is determined by the angles (ψ), the beams per angle (β), the interpolation method employed in determining w_{ij}. Switching of projection dimensions from α to d, where α>d, is periodically carried out by floating the zero components in each projection. The resulting vector serves as a effective descriptor of the image.

Hannad (2016) [5] used texture descriptors of handwritten fragments. The handwritten image is first binarized and non-spanning components are extracted from the image. The components are fragmented into windows/ cells of considerable dimensions. The cells having small writings are eliminated as noise. Textural descriptors such as Local Binary Patterns (LBP), encoding the image using 0s and 1s, Local Ternary Patterns (LTP) positive and negative, using 0s, 1s and -1s for encoding and Local Phase Quantization (LPQ), a transform based approach are used to generate histograms that serve as a descriptor of the particular image. In LBP, pixels in the image are converted into binary runs by comparing the neighboring pixels with the center pixel, resulting in binary 1 if a nearby pixel is greater than the center pixel and binary 0 otherwise. Local Ternary Patterns is similar to LBP but introduces a threshold't' for the center pixel to overcome noise and distortion in the image. Here a binary 0 results if the neighboring pixel is greater than center pixel plus t and binary 0 if the neighboring pixel is between center pixel ± t and -1 otherwise. LTP positive is obtained by replacing -1 by 0 and LTP negative is obtained by replacing -1 by 1. Local Phase Quantization uses Short-Term Fourier Transform (STFT) to encompass phase of the particular window. From the transform containing real and imaginary parts, the binary 1 is placed if sign is positive and binary 0 is placed otherwise. From the generated binary runs using each of the techniques a decimal value is computed and plotted in the histogram which is the unique descriptor of the writer of the handwritten document contained in the image. The computational complexity of histogram generation increases with the number of handwritten samples per writer.

Dasgupta (2016) [6] used Arnold Transform to extract directional features from cursive words in handwritten images. Arnold transform divides images into 4×4 non-overlapping blocks. This process is continued until a desired block size is obtained. Arnold transform is an image scrambling process with a certain period after which the original image comes back. The directional features are extracted using the stroke distributions in the image. The stroke orientations are obtained by subjecting Hough transform over the image. Dimensionality of the features is reduced by using Principal Component Analysis (PCA).

Despite these offline feature extraction techniques, Rodriguez (2016) used word based features, grammatical features, syntactic features, social media and instant messaging features for recognizing online characters or sentences written by writers. These online features are useful if we have digitized texts. In case of the handwritten images, the images should be converted into digitized texts and these features extracted should be used for writer identification.

Venugopal (2017)[7] developed a technique for online mode of writer inference. Object retrieval from handwritten image is done using reduced distortion codebook descriptor replacing Vector of Local Aggregate Descriptor (VLAD). Normalization of features is not done using Min-Max or Z-Score. The normalization is done in such a way to eliminate outliers to an extent, thus increasing performance. Speed, Writing direction, curvature, Vicinity aspect, Vicinity curliness are used as features of online handwritten recognition.

Christlein (2017) [8] proposed a normalized scale invariant feature descriptor called RootSIFT. RootSIFT perform L1 normalization on the feature X, which is then element wise squared followed by L2 normalization. By the normalized feature vectors, RootSIFT descriptors work well for Image recognition. Dimensionality of the SIFT descriptors are reduced by employing principle component analysis (PCA). Orthogonal transformation can be used to reduce correlation among the features. Hence descriptive features are identified. SIFT is quite slow and still subjected to blurs in the image.

Sheng He (2017) [9] put-forth an algorithm for writer inference using curvature free features. The features are extracted using the Local Binary Patterns (LBP) runs on binary and gray scale images. LBP computation is tedious when there is large number of writing fragments for the same writer. Cloud of Line Distribution (COLD) is useful in extracting contours from the handwritten images, thus serving as a feature describing the writing style and writing direction of the writer. These features are not robust enough to uniquely infer the author of a particular handwritten sample.

Ahmad Khan (2017) [11] used Discrete Cosine Transform to obtain features out of image samples. The image represented by 0 to 255 pixels is transformed to cosine space -1 to 1 by centering the image from pixels by 128. Then the DCT coefficients (w_k) are computed for each of the pixels for k=0, 1, 2 … (n-1), where n represents the number of pixels.

$$w_k = \beta_k \sqrt{\frac{2}{n}} \sum_{t=0}^{n-2} \alpha_k \cos(\frac{\pi}{n}\left(t+\frac{1}{2}\right)k) \qquad (2)$$

The DCT coefficients (DCT vectors) are computed for each image. By random selection of vectors of each image, a codebook is formed with the clustering of features leveraging k-Means. About M codebooks are generated from L images where L >> M. Each of the codebook along with the DCT vector of a training image provides a descriptor uniquely identifying the writer of the handwritten document (image). This technique of extracting features is more robust and serves as a better descriptor especially because of the number of codebooks generated to form the descriptor of the particular training image. Codebooks are generated by random selection of features from DCT vectors of each image and then performing K-Means Clustering. DCT takes integer value as input and produces a real valued output, thus requiring an additional quantization step. This opens the space for novelty in feature extraction from handwritten fragments.

III. PREDICTION OF WRITERS

With the descriptors of the handwritten image obtained through feature extraction, there is a need to identify the writer characterized by those descriptors. Hence the training and test images of handwritten documents are subject to feature extraction and the descriptors are extracted from these images in case of offline writer identification. Many learning algorithms are used to predict the writer of a handwritten document characterized by the descriptor. Djeddi (2013) [1] used K-Nearest Neighbors and One-against-all SVM to recognize the writer of the particular handwritten document. Performance analysis of the application is visualized by using ROC curves. The performance is quantified using the Equal-Error-Rate (EER), the point at which the corresponding False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equal. Surinta (2015)[2] also used K-Nearest Neighbors and SVM to identify writers with the given descriptors of the offline handwritten images. The following sub-sections describe various learning techniques.

A. k-Nearest Neighbors (k-NN)

k-NN is a non-parametric algorithm that finds nearest neighbors for each instance of the data point. K-NN has only one parameter k by which the instances are considered as neighbors. The k-NN criterion is employed in handwritten character recognition [2] and also writer identification systems. Some previous studies [1,3] used k-NN algorithm for writer inference using multi-lingual handwritten characters and acquired a good recognition performance of 82.54% on writer inference and verification on English text and 92.06% recognition on Greek text. Hence K-NN is the most commonly used soft computing technique to learn the images that belong to the common class. More precisely it groups the hand-written images into classes of images written by the same writer.

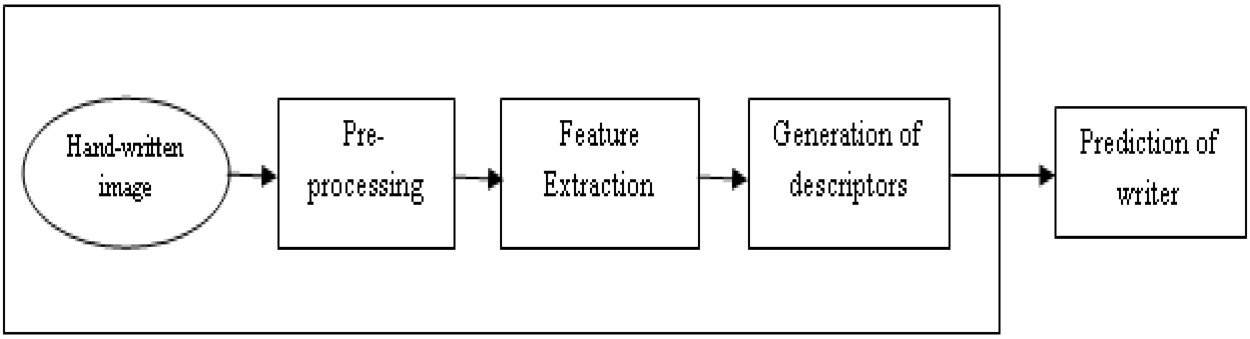

Fig. 2. Methodology of Writer Inference

B. Random Forest (RF) classifier

Instead of taking the entire feature set for performing classification, RF takes a random set of features into account and compounds the samples based on these features into trees. It does not take into account of outliers or noise in the dataset. The random features chosen may also be an outlier Villar-Rodriguez (2016) [10] used RF and SVM for authorship identification and obtained a Figure of Merit of 0.607/0.014 using Random Forest and 0.718/0.001 using SVM.

C. Multiclass Support Vector Machine

Multiclass Support Vector Machines (SVM) is normally used for prediction and classification problems. It constructs an optimal hyper plane that best classifies or categorizes the datasets. The SVM seek to correctly classify the unseen data. In the context of study SVM must classify handwritten documents into appropriate classes of its writers. It overcomes disadvantages of various learning algorithms, such as, local minima and over fitting, among others. Djeddi (2013) have employed the one-against-all SVM and a performance of 87.30% on writer identification in English text and 92.06% identification on Greek text. It is obvious that for the similar dataset, SVM outperforms K-NN in writer identification.

One variation of SVM is the Exemplar-SVM. In Exemplar SVM, the sample under study is treated as positive and all other training samples using which it is to be studied are considered as negative.

D. Principle Component Analysis

The features extracted will usually be large in number. But all of those do not essentially characterize the particular writer. There would be some correlating features. Taking all of those as descriptors is not of interest. Hence there is a need to reduce the dimension and use only unique features from samples and use them as descriptors. Principal Component Analysis is one way of dimensionality reduction which generates the principal factors by computing the correlation between the feature vectors. Dasgupta (2016) used principal component analysis to find the set of linearly uncorrelated features extracted from handwritten fragments using Hough transform.

E. Kernel Discriminant Analysis with Spatial Regression (SR-KDA)

SR-KDA is another method for dimensionality reduction. Unlike PCA, SR-KDA does not require computation of Eigen values, as it solves regularized regression problems. In this case of discriminant analysis, a matrix of dimension n × n is computed from the train vectors, acquired from the code book of feature descriptors (Ahmed Khan, 2016) [11].

In our previous work, we developed a Hybrid algorithm, Hybrid Hierarchical Feature Tree based Authorship Inference (HHFTAI) [12], using Hough Circle transforms, vertical length of black pixels and Speeded-up Robust Features (SURF). HHFTAI is experimented on IAM, CVL and the self-curated Tamil dataset. Nearest neighbor criterion is used to learn the writer. The comparison of various state-of-the-art writer inference techniques is depicted in Table- I.

Table- I: Comparison of various Writer Inference Techniques

Inference Techniques	**Table Column Head**		
	IAM dataset	***CVL dataset***	***Tamil dataset***
Textural Descriptors [5]	89.50%	96.20%	-
Arnold Transforms [6]	95.24% (CENPARMI dataset)	-	-
Discrete Cosine Transoforms [11]	99.60%	71.60%	-
HHFTAI [12]	97.15%	96.20%	97.35%

IV. METHODOLOGY OF WRITER INFERENCE

Authorship identification is typically based on two main methodologies: Extraction of features from handwritten images and prediction of writers using descriptors provided by the generated features. For online recognition, the context also comes out of image processing and deals with extracting stroke lengths, patterns and writing speed. Typically offline writer identification is more crucial in this context.

Fig. 2 shows the overall methodology of authorship identification. Among several techniques used for writer identification as described above such as run length features, Discrete Radon Transform, Hough transform, Arnold transform, Discrete Cosine Transform, the efficient way of extracting features is to apply transformations onto the images under study. Among the transforms, Discrete Cosine Transform is more predominant showing better accuracy of 99.60% with the IAM database. HHFTAI seem to be more robust as it works with better accuracy in three datasets: IAM, CVL and Tamil dataset.

For prediction, the most prevalently used technique is k-Nearest Neighbors and SVM depending on the context of use. The variant of these were also used such as exemplar SVMs. Significant result can be achieved when there are multiple predictors predicting the writer and the final result is obtained by voting on the predictor results. Dimensionality reduction can be carried out using PCA.

V. CONCLUSION

Thus the paper depicted various algoithms used for feature extraction and prediction of writers given a set of descriptors describing the writer of the handwritten document. Writer Identification in addition to these techniques also be characterized by the language in which the it is being applied. It is observed that more prevalently used languages were English, Greek, Arabic, Bangla, Gurmukhi, and Chinese. The reason for the bounds in languages is mainly due to the lack of available standard handwritten datasets. Hence there is also a place for novelty from the perspective of language used for Authorship Identification in addition to feature extraction and prediction techniques. The writer inference task can also be compared as a character recognition task, visualizing the writer of the former, in the place of character in the latter. Hence deep learning algorithms can be employed in writer inference to gain a high degree of accuracy.

REFERENCES

1. C. Djeddi, I. Siddiqi, L. Souici-Meslati, and A. Ennaji, "Text-independent writer recognition using multi-script handwritten texts", *Elsevier Journal of Pattern Recognition*, vol. 34, no. 1, 2013, pp. 1196-1202.
2. O. Surinta, M. F. Karaaba, and L. Schomaker, M. A. Wiering, “Recognition of handwritten characters using local gradient feature descriptors”, *Elsevier Journal of Engineering Applications of Artificial Intelligence*, vol. 45, no.1, 2015, pp. 405-414.
3. S. He, M. Wiering, and L. Schomaker, “Junction detection in handwritten documents and its application to writer identification”, *Elsevier Journal of Pattern Recognition*, vol. 48, no. 12, 2015, pp. 4036–4048.
4. S. Y. Ooia, A. Teohb, Y. Panga, and B. Y. Hiewa, “Image-based handwritten signature verification using hybrid methods of discrete Radon transform, principal component analysis and probabilistic neural network”, *Elsevier Journal of Applied Soft Computing*, vol. 40, no. 1, 2016, pp. 274-282.
5. Y. Hannad, I. Siddiqi, and M. Kettania, Writer identification using texture descriptors of hand written fragments, *Elsevier Journal of Pattern Recognition*, vol. 34, no. 1, 2016, pp. 1196-1202.
6. J. Dasgupta, K. Bhattacharya, and B. Chanda, “A holistic approach for Off-line handwritten cursive word recognition using directional feature based on Arnold transform”, *Elsevier Journal of Pattern Recognition*, vol. 79, no. 1, 2016, pp. 73-79.
7. V. Venugopal and S. Sundaram, “An online writer identification system using regression-based feature normalization and codebook descriptors”, *Elsevier Journal of Expert Systems with Applications*, vol. 72, no. 1, 2017, pp. 196-206.
8. V. Christlein, D. Bernecker, and F. Hönig, “A. Maier, and E. Angelopoulou, Writer Identification Using GMM Supervectors and Exemplar-SVMs”, *Elsevier Journal of Pattern Recognition*, vol. 63, no. 1, 2017, pp. 258-267.
9. S. He, L. Schomaker, “Writer identification using curvature-free features”, *Elsevier Journal of Pattern Recognition*, vol. 63, no. 1, 2017, pp. 451-464.
10. E. Villar-Rodriguez, J. Bilbao, and S. Salcedo-Sanz, “A feature selection method for author identification in interactive communications based on supervised learning and language typicality”, *Elsevier Journal of Engineering Applications of Artificial Intelligence*, vol. 56, No. 1, 2016, pp. 175-184.
11. F. Ahmad Khan, M. A. Tahir, F. Khelifia, A. Bouridane, and R. Almotaeryi, “Robust off-line text independent writer identification using bagged discrete cosine transform features”, Elsevier Journal of Expert Systems with Applications, vol. 71, no. 1, 2017, pp. 404-415.
12. R. R. Subramanian, K. Seshadri, "*Design and Analysis of a Hybrid Hierarchical Feature Tree based Authorship Inference Technique*", Advances in Data and Information Sciences, Springer, vol. 2, no. 1, 2019, pp. 89-104.

AUTHORS PROFILE

R. Raja Subramanian is currently working as an Assistant Professor in Kalasalingam Academy of Research and Education, India. He is qualified in GATE 2016 and UGC NET as Assistant Professor in 2017. He is doing research in the area of fog computing. He has good number of publications in the areas of Image Processing and Computer Vision, Fog Computing, Deep Learning.

Ramalakshmi Ramar is currently working as an Associate Professor in Kalasalingam Academy of Research and Education, India. She has 19 years of teaching experience. She has number of publications in the areas of Networking, Image Processing, Distributed Systems and Security. She has obtained one DST funded project and guiding 8 Ph.D. scholars. She also obtained the Young Scientist Award from the Tamil Nadu Government.