Jurnal
Teknologi

Full Paper

REVIEW OF OFFLINE TEXT INDEPENDENT WRITER IDENTIFICATION TECHNIQUES

Shehzad Khalid[a], Uzma Naqv[b], Sana Shokat[b*]

[a]Department of Computer Engineering, Bahria University, Islamabad, Pakistan
[b]Department of Computer Sciences & IT, University Azad Jammu and Kashmir, Muzaffraabad

Article history
Received
25 October 2015
Received in revised form
14 December 2015
Accepted
9 February 2016

*Corresponding author
snagul@yahoo.com

Abstract

This paper reviews various text independent writer identification techniques through offline documents. Different features extraction methods are discussed. Classification approaches that are mainly used for identification by the researchers and verification by different groups and individuals are presented. Identification rates achieved by the reviewed papers are tabulated and analyzed. A survey of different databases used in the reviewed papers is performed. Application of writer identification in different language domains is also discussed. Future directions for the automated writer identification are presented in the end.

Keywords: Writer identification, offline document, classification

1.0 INTRODUCTION

Writer recognition is a process to identify the writer of the questioned document. Writer recognition process comprises of writer identification and writer verification stages. Writer identification involves searching a handwritten samples of unidentified author with samples of N identified authors in a database. The writer identification process comprised of preprocessing, feature extraction and classification standard steps [1] as shown in Figure 1.

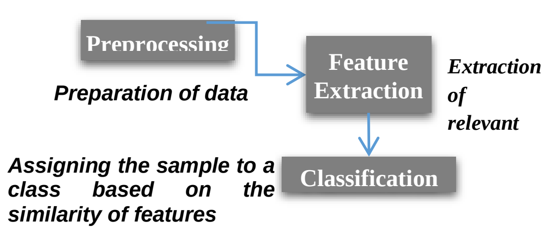

Figure 1 Steps of Writer Identification process

Writer identification classified into Offline and online based on the method of writing. In online writer identification detail information such as speed of writing, angle or pressure while in offline features associated with characters, words, lines or paragraphs are utilized. Over the past two decades, automatic offline writer identification has enjoyed renewed interest. [2]

Writer identification task is performed through text dependent or text independent methods. In text-dependent methods same content must be written by the writers to be compared however text-independent methods do not require the content to be same. Text independent got a wider applicability but do not obtain the same high accuracy as text-dependent methods do [3].

Writer recognition systems use global features such as texture, curvature and slant features [4], [5] and also local features such as graphemes, allographs and connected components to identify the writers. Extraction of local features require segmentation and this approach is applied in [6-10].

2.0 WRITER IDENTIFICATION: STATE OF THE ART

The work done in the field of writer recognition until 1989 was presented in a survey paper [11]. Later, work till 1993 was published in [12].

Some feature extraction methods are applied to fixed length datasets in which writers are asked to copy a particular text to a given number of times to produce the samples as in [6] and [13]. This methodology is normally applied to characters or words. Others are variable in sample length.

3.0 FEATURE EXTRACTION METHODS

Global features are extracted at micro and macro level of resolution. As described by [2] at macro level, thirteen global feature measures are extracted which are pen pressure, writing movement, stroke formation, average line height and average slant per line, stroke width, and average word gap measures. On the other hand micro features extracted from the allograph, or character. Gradient, structural and concavity (GSC) features developed for character recognition were calculated at micro level [14].

Local features are calculated directly and provide adequate information about the basic composition and shape of a character. Extraction of local features require segmentation and this approach is applied in [6-10].

To exploit the information in the unlabeled data where the writer is unknown a semi supervised structural learning framework presented in [15] that used the multitask learning to provide low dimensional common sub structure.

Textual features extracted from image or when signal applied to the document without using writing information [16] such as wavelets, filters and autocorrelation. Textural features can be obtained through the multi-channel Gabor filtering [17]and the grey scale co-occurrence matrix (GSCM) [18]methods. In [8] used textual based Information Retrieval model for the writer identification that allows the use of a specific feature space on the basis of feature frequencies. They used a segmentation procedure of handwritten components followed by a cluster analysis to derive a set of features. [19] utilized texture based features for writer verification by using GLCM (Grey Level Co-occurrence Matrix) .They proved that by combining classifiers better results can be achieved based likelihood analysis producing receiver operating characteristics (ROC) curves of the classifiers.

HMM presented [20] that used sample data to train for every individual writing style at character and word level. Drawbacks of HMM are covered by Gaussian Mixture Models (GMMs) used in HMM training [21]that eliminated the training need and character/word model requirements by attaining almost 98% accuracy.

In [22] they applied only vertical scaling. The text is provided to HMM to decide which writer has written an unknown text, result with maximum value is expected to be the writer and reported 97% accuracy for 100 writers. But its applicability is limited due to training requirements.

4.0 FEATURE TYPES

Graphemes introduced in [23] that are writer invariants based on redundant individual patterns of a writing. Samples did not analyze for writer verification. Extended work presented in [24] using the same features in addition an information retrieval paradigm is applied to describe ,compare the handwritten query to each sample in the database and grapheme concatenation.

The grapheme codebooks used by [25]), and later [26] and [27]propose a sub-grapheme level of codebook and reference base is constructed of smaller stroke fragments. Authors in [27] proposed small low-level stroke fragment Codebooks by using the small windows on the ink trace and experimented with different window sizes. Another segmentation method using codebook was proposed by [28] performed better than the usual ink trace minima segmentation.

After segmentation sequential clustering algorithm [29] is used to group morphologically similar ones where two graphemes are compared using a correlation similarity measure. Later the authors in [8, 24, 30] used the cluster of all the graphemes of the database therefore created a common feature space and invariants collected are used as binary features.

CS-UMD [31] methodology used K-adjacent segment (KAS) features to model a user's handwriting in a bag-of features (BOF) framework.

5.0 FEATURE GROUPING.

Features can be combined in different combinations to achieve better results as in [9, 27, 32-34]. Authors in [9] extracted two visual attributes orientation and curvature had been utilized to distinguish handwritings. And then the effect of the combining both of the attributes had been observed in characterizing the writer. Chain code method and polygons are used to represent features. They got 97% identification rate for IAM and 93% identification rate for RIMES and 95% identification rate for both.

6.0 FEATURE SELECTION.

Feature selection that is to choose subsets of features from original dataset in order to improve performance and data reduction [33, 35-37]. In [38] proposed writer identification system optimization that reduces the identification process search space. They used edgehinge and run-length texture based features. With their proposed methodology they reported accuracy 92.5% and 99.5% for the IFN/ENIT database and GRDS database respectively and 93.3% for the union of databases.

Authors got top-10 result of 95.4% for 250 writer database by applying data reduction algorithms such as principal component analysis [PCA], linear discriminant analysis [LDA], multiple discriminant analysis [MDA], multidimensional scaling [MDS], and forward=backward feature selection algorithm) [33].

7.0 CLASSIFIERS

Different techniques as shown in table 1 are available that proved to effectively measure the distances such as Euclidean distance or Hamming [16], and Chi-square. Researchers [39] used techniques to define the distance between two points differently, weighting some dimensions more significantly than others [1]. Authors [25] finds that Hamming performs best, but uses the chi-square distance.

After metric selection classification is the next step. In classification stage unknown object is assigned to an existing known group. There are several common classifiers available, such as Nearest Neighbor (NN), Neural Networks, Decision Trees and Support Vector Machines (SVMs). SVM is used for classification of 93 writers [15]. Artificial neural networks is used in the study [6] for classification. NN applied in [9, 25, 37, 40] is a very simple and widely used.

Authors in [42] presented a framework to utilize the sampling for classification. They used Bayesian classification to assign sample weights in accordance to the sample importance.

Table 1 Various text independent writer identification techniques

Authors	Features	Classifiers	Accuracy (%)	Data Set-Writers(sample)
(Srihari, Cha, Arora, & Lee) [6]	Macro and micro features	Neural net.	94.0	CEDAR-100(3)
(Schlapbach & Bunke) [22]	Global and local features	Hidden Markov Models(HMM)	98.4	IAM-100(5)
(BENSEFIA, PAQUET, & HEUTTE) [8]	Graphemes	Hypothesis Test	86.0	IAM-150(2)
(M.Bulacu & L.Schomaker) [41]	directional, grapheme, and PDFs	Distances	89.0 83 87	IAM-650(2) Firemaker-250 Large-900
(Siddiqi & Vincent) [9]	Global , local & Polygon	Distances	97 93 95	IAM-650 RIMES-375 IAM+RIMES(1025)

8.0 DATABASES

Databases varies in text dependency, language, number of writers resolution, samples per writer available and type of sample such as paragraph ,lines or characters. There are several databases available which considered authentic among the research community. Some of them (text type) are listed in the Table 2. They are as IAM[43], CEDAR [6], RIMES[44], Firemaker [7], UniPen, or IFN/ENIT[45]. Usually, the authors have chosen different datasets from these databases as shown in table 2.

Some of them i.e. IAM and RIMES are annotated databases about the writer identity, the ground truth text and the segmentation at a line, sentence, word and character levels.

Table 2 Databases Used for Writer identification in Literature

Data Set	Language	No. of writers	Sample Size
IAM [43]	English	650	Variable(1-59)
RIMES[44]	French	1600	5
CEDAR[46]	English	1000	3
Firemaker([7]	Dutch	252	4
TriGraph Slant[47]	Dutch	47	4 pages/writer
Unipen [48]	Various	215	2 pages/writer

9.0 CONCLUSION

In this paper we presented the state of the art in writer identification, types of features, the feature extraction approaches, the classifiers and the databases used. The literature was grouped by the research work publications based on similarities in used features and classifiers. That shows the improvement done by the researchers. To make the comparison of different research work features, the classifiers, the databases used, the best identification rates of each publication, the number of writers and the year of publication are tabulated. Tabulation was included for the used databases, the number of writers, samples, etc. This specifies the large number of publications in this field and increasing number of researchers working in this area.

References

[1] Zhu, Y., Wang, T. T., & Yunhong. 2000. Biometric Personal Identification Based on Handwriting. *Pattern Recognition Proceedings 15th International Conference.* Barcelona. 3-7 September 2000. 2: 797-800

[2] Awaida, S. M., & Mahmoud, S. A. 2012. State Of The Art In Off-Line Writer Identification Of Handwritten Text And Survey Of Writer Identification of Arabic text. *Educational Research and Reviews*. 7(20):445-463.

[3] Sreeraj.M, & Idicula, S. M. 201. A Survey on Writer Identification Schemes. *International Journal of Computer Applications.* 25(12): 23-33.

[4] Bulacu, M., & Schomaker, L. 2007. Text-Independent Writer Identification And Verification Using Textural And Allographic Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 29: 701-717.

[5] Hochberg, J., Kelly, P., Thomas, T., & Kerns, L. 1997. Automatic Script Identification From Document Images Using Cluster-Based Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 19: 176-181.

[6] Srihari, S. N., Cha, S. H., Arora, H., & Lee, S. 2002. Individuality of handwriting. *Journal of Forensic Sciences*. 1-17.

[7] Bulacu, M., Schomaker, L., & Vuurpijl, L. 2003. Writer Identification Using Edge-Based Directional Features. *7th International Conference on Document Analysis and Recognition.* Edinburgh, Scotland. 3-6 August 2003. 937.

[8] Bensefia, A., PaquetT, T., & Heutte, L. 2005. Handwritten Document Analysis for Automatic Writer Recognition. *Electronic Letters on Computer Vision and Image Analysis*. 5(2): 72-86.

[9] Siddiqi, I., & Vincent, N. 2010. Text Independent Writer Recognition Using Redundant Writing Patterns With Contour-Based Orientation And Curvature Features. *Pattern Recognition.* 43(11): 3853–3865.

[10] Kirli, O., & Gulmezoglu, M. 2011. Automatic Writer Identification From Text Line Images. *International Journal on Document Analysis and Recognition.* 15(2):1-15.

[11] R.Plamondon, & G.Lorette. 1989. Automatic Signature Verification And Writer Identification-The State Of The Art. *Pattern Recognition*. 22(2):107-131.

[12] Leclerc, F., & Plamondon, R. 1994. Automatic Signature Verification: The State Of The Art-1989-1993. Intl. *Journal of Pattern Recognition and Artificial Intelligence.* 643-660.

[13] Rahmati, F. N. 2007. A New Method for Writer Identification and Verification Based on Farsi/Arabic Handwritten Texts. *ICDAR09*. Parana. 23-26 September 2007. 2: 829–833.

[14] Srihari, S., S.H.Cha, Arora, H., & Lee, S. 2001. Individuality of Handwriting : A Validity Study. *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference.* Seattle, WA. 10-13 September 2001. 106-109.

[15] Porwal, U., Ramaiah, C., Shivram, A., & Govindaraju, V. 2012. Structural Learning for Writer Identification in Offline Handwritingw. *International Conference on Frontiers in Handwriting Recognition*. Bari. 18-20 September 2012. 415-420.

[16] Shahabi, F., & Rahmati, M. 2009. A New Method for Writer Identification of Handwritten Farsi Documents. *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference.* Barcelona. 26-29 July 2009. 426-430.

[17] Peake, G. S., & Tan, T. N. 1997. Script and language identification from document images. *British Machine Vision Conference*. San Juan. 20 Jun 1997. 610-619.

[18] Tan, T. 1996. Written Language Recognition Based On Texture Analysis. *International Conference on Image Processing*. Lausanne. 16-19 September 1996. 185-188.

[19] Hanusiak, R., Oliveira, L. S., Justino, E., & Sabourin, R. 2012. Writer Verification Using Texture-Based Features. *International Journal on Document Analysis and Recognition*. 15(3):213-226

[20] A.Schlapbach, & H.Bunke. 2004. Using HMM Based Recognizers For Writer Identification And Verification. *Proceedings of the Ninth International Workshop on Frontiers in Handwriting Recognition.* Kokubunji, Tokyo, Japan. 26-29 October 2004. 167-174.

[21] Schlapbach, A., & Bunke, H. 2006. Off-line Writer Identification Using Gaussian Mixture Models. *18th International Conference for patteren Recognition.* Hong Kong, China. 20-24 August 2006. 992–995.

[22] Schlapbach, A., & Bunke, H. 2007. A writer identification and verification system using HMM based recognizers. Pattern Analysis and Applications (PAA). 10(1): 33-43.

[23] Bensefia, A., Nosary, A., Paquet, T., & Heutte, L. 2002. Writer Identification by Writer's Invariants. *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop.* Ontaria, Canada. 6-8 August 2002. 274-279.

[24] Bensefia, A., T. Paquet, L., & Heutte. 2003. Information Retrieval Based Writer Identification. *Proceedings* of *the Seventh International Conference* on *Document* Analysis and *Recognition* (ICDAR *2003*). Edinburgh, UK. 3-6 August 2003. 946-95.

[25] Schomaker, L., & Bulacu, M. 2004. Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Upper-CaseWestern Script. IEEE Transactions on Pattern Analysis and Machine Intelligence. 26(6): 787-798.

[26] Seropian, A., Grimaldi, M., & Vincent, N. 2003. Writer Identification Based On The Fractal Construction Of A Reference Base. *7th International Conference on Document Analysis and Recognition*. Edinburgh, UK. 3-6 August 2003. 1163-1167.

[27] Siddiqi, I., & Vincent, N. 2007. Writer Identification in Handwritten Documents. *9th International Conference on Document Analysis and Recognition*. Parana. 23-26 September 2007. 108-112.

[28] Ghiasi, G., & Safabakhsh, R. 2010. An Efficient Method for Offline Text Independent Writer Identification. *2010 International Conference on Pattern Recognition.* Istanbul, Turkey. 23-26 August 2010. 1245-1248.

[29] Friedman, M., & Kandel, A. 1999. Distance Function. *Introduction to Pattern Recognition*. 70-73

[30] A. P. T. &. H. L. Bensefia, "A Writer Identification And Verification System," *Pattern Recognition Letters*. 2005b.

[31] Jain, R., & Doermann, D. 2011. Offline Writer Identification using K-Adjacent Segments. *2011 International Conference on Document Analysis and Recognition*. Beijing, China. 18-21 September 2011. 769-773.

[32] Gazzah, S., & Amara, N. B. 2008. Neural Networks and Support Vector Machines Classifiers for Writer Identification Using Arabic Script. *The International Arab Journal of Information Technology.* 5(1): 92-101.

[33] AWAIDA, S. M., & MAHMOUD, S. A. 2013. Writer Identification of Arabic Text Using Statistical and Structural Features. *Cybernetics and Systems: An International Journal.* 44(1): 57-76.

[34] Helli, B., & Moghaddam, M. E. 2010. , "A Text-Independent Persian Writer Identification Based On Feature Relation Graph (FRG*). Pattern Recognition*. 43(6):2199-2209.

[35] Guyon, I., & Elisseef, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research.* 3: 1157-1182.

[36] Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. 2006. *Feature Extraction: Foundations and Applications*. Netherlands: Springer.

[37] Siddiqi, I., Khurshid, K., & Vincent, N. 2011. Feature Relevance Analysis for Writer Identification. *Document Recognition and Retrieval XVIII.* San Francisco, CA, USA. 23-27 January 2011. 78740F-78740F.

[38] Djeddi, C., Siddiqi, I., Souici-Meslati, L., & Ennaji, A. 2012. Multi-script Writer Identification Optimized With Retrieval Mechanism. *2012 International Conference on Frontiers in Handwriting Recognition*. Bari. 18-20 September 2012. 507-512.

[39] Schomaker, L., & Bulacu, M. 2004. Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Upper-CaseWestern Script. IEEE Transactions on Pattern Analysis and Machine Intelligence. 26(6): 787-798.

[40] Brink, A., Smit, J., Bulacu, M., & Schomaker, L. 2012. Writer identification using directional ink-trace width measurements. *Pattern Recognition*. 45: 162–171.

[41] Bulacu, M., & Schomaker, L. 2007. Text-Independent Writer Identification And Verification Using Textural And Allographic Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 29: 701-717.

[42] Wanga, Y., Fub, Q., Dinga, X., & Liua, C. 2015. Importance Sampling Based Discriminative Learning For Large Scale Offline Handwritten Chinese Character Recognition. *Pattern Recognition.* 48(4): 1225-1234.

[43] Marti, U. V., & H.Bunke. 1999. A full English Sentence Database For Off-Line Handwriting. *5th International Conference on Document Analysis and Recognition*. Bangalore. 20-22 September 1999. 705-708.

[44] Grosicki, E., Carré, M., Brodin, J.-M., & Geoffrois, E. 2008. Rimes Evaluation Campaign For Handwritten Mail Processing. *ICFHR'08:11th Int'l Conference on Frontiers in Handwriting Recognition.* Quebec, Canada. 19-21 August 2008. 1-5.

[45] Pechwitz, M., Maddouri, S. S., M¨argner, V., Ellouze, N., & Amiri, H. 2002. IFN/ENIT - database of handwritten Arabic words. *CIFED 2002*. 129-136.

[46] Cha, S., & Srihari, S. 2000. Assessing the Authorship Confidence of Handwritten Items. *Fifth IEEE Workshop on Applications of Computer Vision*. Palm Springs, California. 4-6 December 2000. 42-47.

[47] Brink, A., Niels, R., Batenburg, R. v., Heuvel, C. v., & Schomaker, L. 2010. Towards Robust Writer Verification by Correcting Unnatural Slant. *Pattern Recognition*. 32(3):449-457.

[48] Guyon, I., Schomaker, L., Plamondon, R., Liberman, M., & Janet, S. 1994. UNIPEN Project Of On-Line Data Exchange And Recognizer Benchmarks. *12th IAPR International Conference on Pattern Recognition.* Jerusalem. 9-23 October 1994. 2: 29-33.