International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Hybrid Feature Extraction and Multiclass SVM Based Ancient Character Recognition

P.Balasubrmanie, E.K.Vellingiriraj

Abstract: ***Ancient character recognition is the most difficult task due to their different formats and less frequent knowledge about the ancient characters. This is performed in the previous research work namely Shape and Size aware Character Recognition and Conversion System (SSCR-CS). This research work proved better character recognition outcome. However this research work might be reduced in performance with lesser detailed information in the images. This is resolved in the proposed research work by introducing the most recent techniques for the character recognition outcome. This is attained by introducing the method namely Hybrid Feature Extraction and Multiclass SVM based recognition method (HFE-MCSVM). In this research work, initially image pre-processing is performed by using Gabor filter. After pre-processing segmentation of characters is performed by using overlapped character segmentation method. After segmentation character recognition is done by introducing the method namely hybrid feature extraction with Multiclass SVM classification approach. The overall assessment of the research work is done in the matlab simulation environment and then it is proved the proposed HFE-MCSVM shows better performance.***

Keywords: Ancient character recognition, feature extraction, segmentation of characters, Hybrid multiclass SVM

I. INTRODUCTION

Ancient character acknowledgment is the most focused research is this present reality condition [1]. Optical character acknowledgment (OCR) is a useful use of cutting edge image preparing and design acknowledgment improvements [2]. Employments of OCR incorporate computerized report filing, printed content hunt and mechanized structure handling. Current correspondence offices could permit expansive and open conveyance of immense libraries of books, papers, magazines and a wide range of printed media, if quality, financially savvy OCR strategies are accessible for mass digitizing [3]. While present day printed content can be perceived in all respects precisely with monetarily accessible programming, performing OCR on increasingly intriguing material, (for example, gothic textual styles, antiquated typesets and penmanship) is right now and observably less fruitful [4].

Existing advanced libraries, containing computerized accumulations of old and important manually written and printed archives and books dating up to the second 50% of XIXth century, demonstrate a degree of intuitiveness still very low [5].

For these particular computerized substance, in reality, has not been at this point conceivable to create optical-advanced acknowledgment frameworks as well as content acknowledgment of virtual pages, ready to give a proficient ordering of databases content either effectively available or to comprise over the web 2.0 [6]. None of the most recent and most significant ventures of computerized libraries right now accessible on the web 2.0 has openness and ease of use includes that enable clients to see the content substance of the imitated advanced articles without looking over them through in full [7]. Barring regular recording research (creator, title, discharge notes), in these databases it is beyond the realm of imagination to expect to build up any ordering that permits inside and out examinations dependent on the investigation of the repeat of words, deduction about various writings, and so forth.

This trouble emerges from the idea of the curios being referred to. The intricacy and disparity of antiquated original copy spellings, even those paleographic increasingly direct and standard [8]; the sort of old inks utilized; the out of date quality of the materials, by and large with harms brought about by natural or biochemical elements, mechanical mishaps and human inconsiderateness: every one of these components have so far counteracted all endeavors to go past the basic generation of these computerized relics. Neither the flows OCR, ICR and IWR accessible available can be connected to take care of the issue of content acknowledgment in old archives.

In the event that this circumstance appears to be practically evident on account of original copies, as a result of their inclination, it ought to be less justifiable for the printed books [9]. Rather, notwithstanding for this sort of curio, specifically for books created by handprinting, the circumstance is fundamentally the same as that of the compositions. The issues, truth be told, are not unique, regardless of whether they influence to a lesser degree. The methods of sythesis of the printing plates, the inks utilized, the arrangement of stamps inside words, and of the words inside lines, the distinctive realistic text styles illustrative of specific letters [10], when contrasted with those generally utilized (eg., the "s" spoken to by a printing textual style

Manuscript published on 30 September 2019
* Correspondence Author
P.Balasubrmanie*, Professor, Department of Computer Science & Engineering Kongu Engineering College, Perundurai – 638 052, Erode, Tamil Nadu, India. Email: balu_p@kongu.ac.in
E.K.Vellingiriraj, Former Assistant Professor Department of Computer Science & Engineering Kongu Engineering College, Perundurai – 638 052, Erode, Tamil Nadu, India. Email: balu_p@kongu.ac.in

fundamentally the same as "f"), diverse semantic shows, the different commotion of the images (foundation clamor brought about by the push on the turn around side page [11], smears and breakage of stamps, ink stains and some other changed reason because of time and men) are on the whole factors that, today, disappoint any endeavor to record the substance of computerized images of antiquated materials through utilization of acknowledgment frameworks with fulfilling results.

II. RELATED WORKS

Raghupathy et al [12] utilized the best in class CNN in perceiving written by hand Tamil characters in disconnected mode. CNNs vary from customary methodology of Handwritten Tamil Character Recognition (HTCR) in separating the features naturally. We have utilized a segregated written by hand Tamil character dataset created by HP Labs India. We have built up a CNN model without any preparation via preparing the model with the Tamil characters in disconnected mode and have accomplished great acknowledgment results on both the preparation and testing datasets.

Panyam et al [13] misused an exceptional 3D include (profundity of space) which is relative to the weight connected by the scriber by then. This 3D feature is acquired at every one of the pixel purpose of a Telugu palm leaf character. In this work two dimensional Discrete wavelet change (2-D DWT), two dimensional quick Fourier change (2-D FFT) and two dimensional discrete cosine change (2-D DCT) are utilized for feature extraction. The 3D feature alongside the proposed two level change based strategy gets better acknowledgment precision.

Sadanand et al [14] depicted proficiency of Zernike Complex minutes and Zernike minutes with various Zoning designs for disconnected acknowledgment of manually written 'MODI' characters. Each character was separated in six zoning designs with 37 zones. Geometrical shapes were utilized to make zoning designs. The work was brought about 94.92% right acknowledgment rate was accomplished by utilizing Zernike minutes and 94.78% by utilizing Zernike complex minutes with coordinated methodology for heterogeneous zones.

Bhunia et al [15] proposed Indic cross language system misuses a huge asset of dataset for preparing and uses it for perceiving and spotting content of other objective contents where adequate measure of preparing information isn't accessible. Since, Indic contents are for the most part written in 3 zones, to be specific, upper, center and lower, we utilize zone-wise character (or segment) mapping for proficient learning reason. The presentation of our cross-language structure relies upon the degree of closeness between the source and target contents.

Lakshmi et al [16] managed the distinguishing proof of Telugu Palm leaf characters by securing an extra 3D include on palm leaves. The foundation of these original copies is indistinguishable from the compositions on them. Expelling foundation from such contents is a repetitive errand. This is accomplished with the 3D include profundity in the present work. With the assistance of this 3D include, an improved grouping rate is additionally accomplished.

Premaratne et al [17] proposed a novel technique that investigates the vocabulary in relationship with the shrouded Markov models to improve the rate of precision of the perceived content. The proposed technique could without much of a stretch be reached out with minor changes to other adjustment based contents comprising of befuddling characters. The word-level precision which was at 81.5% is improved to 88.5% by the proposed advancement calculation.

Likforman-Sulem et al [18] depicted an information based framework that causes copyists to confirm Hebrew original copies for which precise laws of calligraphy have been given to the recorders. Paleographic mastery is additionally incorporated into request to portray the size of the report and the composition. At the point when utilized for validation purposes, the framework abbreviates the errand of the recorder by pointing out the pieces of the archive or the characters where issues emerge for the machine. The recorder will at that point work on a confined piece of the archive and choose whether it must be redressed or not.

Choudhary et al [19] depicted the methodology for improvement an enormous volume of Urdu manually written content images Corpus on Urdu language. To make the database accessible in enormous field of Natural Language Processing we comment on database for each image and partner a XML based ground-truth Meta data to make it PC good as an etymological asset.

III. ANCIENT CHARACTER RECOGNITION USING HYBRID CLASSIFIER

In this research work, initially image pre-processing is performed by using Gabor filter. After pre-processing segmentation of characters is performed by using overlapped character segmentation method. After segmentation character recognition is done by introducing the method namely hybrid feature extraction with Multiclass SVM classification approach.

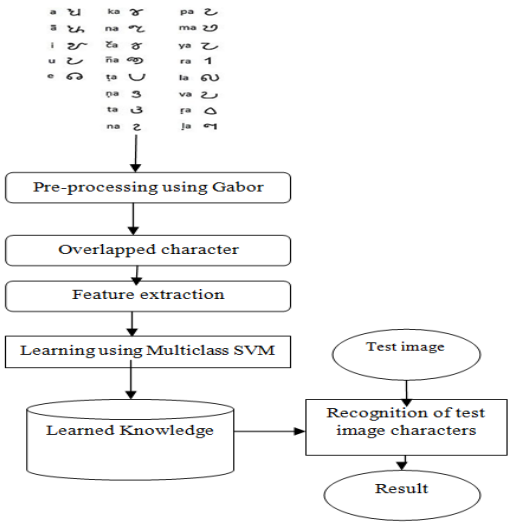

Figure 1. Overview of the proposed research work

A. Preprocessing Using Gabor Filter

Channels are utilized to Remove Noise, Sharpen Contrast, Detect Edges and Feature Contours. Channel is partitioned into Linear and Nonlinear. Direct channel has Convolution portrayal in network and Non-Linear with Thresholding and Image Equalization Gabor Filters are utilized in wide applications for Pattern Analysis, Facial Recognition, Iris Recognition, Optical Character Recognition, Fingerprint Recognition as a result of its main considerations like Computational Properties and Biological Relevance. Gabor Filter is a linear channel utilized for Edge Enhancement. It functions as a Band pass channel for the nearby Spatial recurrence appropriation, accomplishing an ideal goals in both Spatial and Frequency space. The processing flow of gabor filter is shown in the following diagram 2.

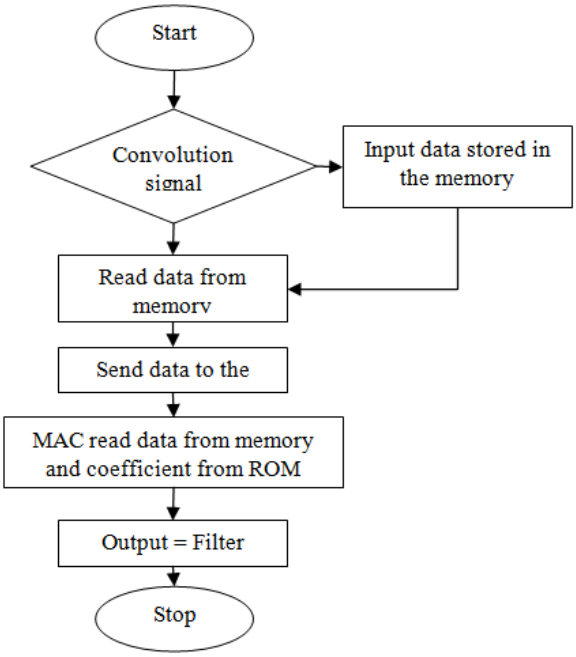

Figure 2. Processing flow of Gabour filter

Gabor channel Frequency and Orientation is like that of human visual framework. Its motivation reaction is Sinusoidal wave increased by Gaussian capacity. This channel has duplication Convolution property and different changes, image property, administrators, frequencies and features in recognizing edges. The channel follows the district esteem and is delicate to collinear and extended sections. The bit of leeway is probability of making channels which are specific for Orientation. Gabor channels with 2D are utilized to concentrate include at each pixel of information image. Short Gabor channels are utilized for Space invariant force images and bended line follows. The long Gabor channel for dark level qualities and straight line follows. The channel can be seen as sinusoidal plane of part recurrence and direction, balanced by a Gaussian envelope.

The parameter of Gabor channel is its recurrence, standard deviation and direction. The channel is structured by number of enlargements and turns, and when it is convolved with sign is called as Gabor space Its duplication convolution property the Fourier change of Gabor channel reaction is the convolution of the Fourier change of consonant capacity and Fourier change of Gaussian capacity. The parameters of the Gabor channel are calibrated with various direction and wavelength. The Gabor channel capacity is spoken to as

$$g(x, y;\ \lambda, \theta, \Phi, \vartheta, \gamma)$$

where λ is the Wavelength of the sinusoidal factor, θ is the Orientation of the ordinary to stripes of Gabor work, Φ is the Phase counterbalanced, Φ is the Standard Deviation of Gaussian envelope, γ is the Spatial Ratio and determine the circularly the help of the gabor work. The parameter of the channel is changed for various Wavelength and Orientation the improvement of edges is come to at a specific Orientation with its diverse parameter esteems. This yield delivered helps in a productive method for pre-processing image for any further advanced image handling work.

B. Character Segmentation Using Overlapped Method

For troublesome inclined words, incline amendment methods don't yield great outcomes. In like manner, to fragment evenly covered characters, we presented a basic idea of word center zone stature. Center zone is the zone that lies between the lower gauge and upper benchmark of the word. Again a basic and quick method is proposed to ascertain lower and upper baselines. From upper left point, check number of white pixels push by column. Compute the normal of those lines for which there is an adjustment in tally among present and going before column until the main huge change happens. A normal column speaks to the upper gauge. A similar strategy is received for the lower benchmark yet from base to top. Proposed division calculation for old characters is displayed as pursue.

Stage 1. Take input image

Stage 2. Perform pre-handling.

Stage 3. Decide center zone tallness.

Stage 4. In the center zone, compute the aggregate of closer view pixels (white pixels) for every segment. Spare those segments as competitor portion segment (CSC) for which entirety is 0 or 1.

Stage 5. By the past advance, we have more applicant division segments than really required. A limit is chosen exactly from up-and-comer fragment sections to turn out with right portion segments. A limit is a steady worth that is inferred after various tests to stay away from over-division.

Because of the straightforwardness of the proposed division strategy, it is exceptionally quick and performs well in the vast majority of the cases. For a couple of characters, for example, w, u, v, m, n and so on over division happens and this system neglects to discover precise character limits. Be that as it may, this issue is fathomed heuristically by choosing an edge and hard stroke features.

C. Hybrid Feature Extraction

SVM library is running under WEKA device is utilized in the proposed trial. WEKA is all around utilized instrument in the AI field. SVM is incorporated programming for help vector order and relapse. Two class (parallel classifier) SVM is connected to multiclass character acknowledgment issue utilizing one versus all strategy.

SVM as the classifier and polynomial capacity as the bit sort have been utilized. The estimations of different parameters have been set to their default esteems. The SVM is prepared with the preparation tests from image dataset. The classifier works in two stages: preparing and testing. In the wake of pre-processing and feature extraction preparing is finished by taking the element vectors which are put away in lattices structure. Mean ($\bar{x}$) computes the example normal. For vectors, mean is the mean estimation of the components in vector x. For grids, mean is a line vector containing the mean estimation of every section. Here the line and segment astute mean of the parallel character image is determined and besides the all out mean is gotten. i=row, j=col. The standard deviation is determined as

$$\bar{X} = \frac{1}{N \times M} \sum_{i=1}^{M} \sum_{j=1}^{N} P(i,j)$$

$$S = \frac{1}{N \times M} \sum_{i=1}^{M} \sum_{j=1}^{N} \bar{X} - P(i,j)$$

N is the quantity of components in the example. The mean ($\bar{x}$), and deviation (s) values are utilized as features. The help vector esteems are determined utilizing these two features and plotted in hyperplane. The two capabilities are put away in feature library f1L, f2L. The element library f1L, f2L had the normal estimation of features mean and standard deviation for 50 preparing images of all out ten I. e. 0 to 9 digits. Preparing image features are put away in feature library (f1L, f2L). Features of test images are in (f1test, f2test). Euclidian Distance is determined

$$D = [(f_{1test} - f_{1L})^2 + (f_{2test} - f_{2L})^2]^{\frac{1}{2}}$$

A image is given as information. Its histogram is plotted. The mean and standard deviation from the histogram is separated. It is contrasted and the purposes of SVM plane by computing Euclidian's separation as to on which side of the plane, the fact of the matter are in, decides the kind of class. Bolster vector machine is basically the ordered technique that performs grouping task by building hyper plane in multi dimensional space that different instances of various class marks. SVM underpins both relapse and characterization errands and can deal with various persistent and absolute factors. To build ideal hyper plane, SVM utilizes an iterative preparing calculation, which is utilized to limit a blunder work .The help vector machine classifier is enhancing a mistake work that limits the misclassification on the preparation set.

IV. RESULTS AND DISCUSSION

The investigations are led in Matlab reenactment condition. This matlab is a programming language. This is for the most part utilized for the variety of numerical and figuring applications. The interface pursues a language that is demonstrated to look a great deal, for example, the documentation use in straight polynomial math. In this exploration work, precise character acknowledgment is finished by presenting the novel structure to be specific Hybrid Feature Extraction and Multiclass SVM based acknowledgment technique (HFE-MCSVM). HFE-MCSVM system is actualized in the matlab reproduction condition whose exhibition is estimated by contrasting it and our past research strategies to be specific SSCR-CS, BC-PCS procedure, HMNL-PRS and furthermore with the current models to be specific iterative logical demonstrating, and the FNN with fluffy c-implies dependent on character acknowledgment approach.

Figure 3. Input image

Figure 4. Pre-processed image

Figure 5. Recognized output

The methodology is differentiated by the metrics like the precision, recall, f-measure and classification accuracy.

A. Accuracy

In the images the weighted level of characters is accurately divided by the estimation exactness. It can speak to as pursues:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \times 100$$

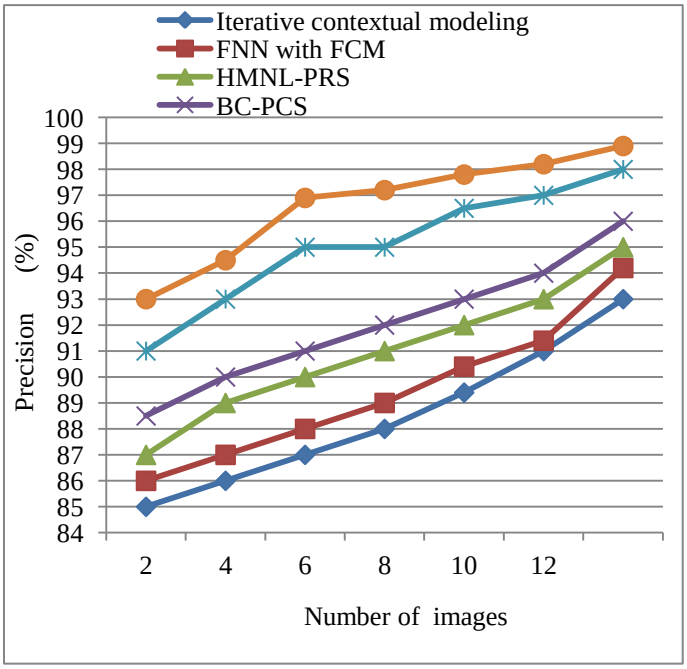

Figure 6. Accuracy comparison

The correlation of an iterative relevant demonstrating, FNN with fluffy c-implies, HMNL-PRS, BC-PCS, SSCR-CS and HFE-MCSVM technique is spoken to in Figure 6 as for exactness. In X hub the quantity of images is plotted and the exactness is plotted in Y hub. This correlation assessment demonstrates that the proposed research strategy HFE-MCSVM procedure is shown the high exactness brings about terms of precise acknowledgment of antiquated tamil characters.

B. Precision

The calculation of exactness or quality is known as Precision, in which the calculation of fulfillment or amount is known as review. Furthermore, the high accuracy demonstrates that the procedures returned fundamentally increasingly related outcomes. The accompanying condition is utilized to assess the exactness.

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive+False positive}}$$

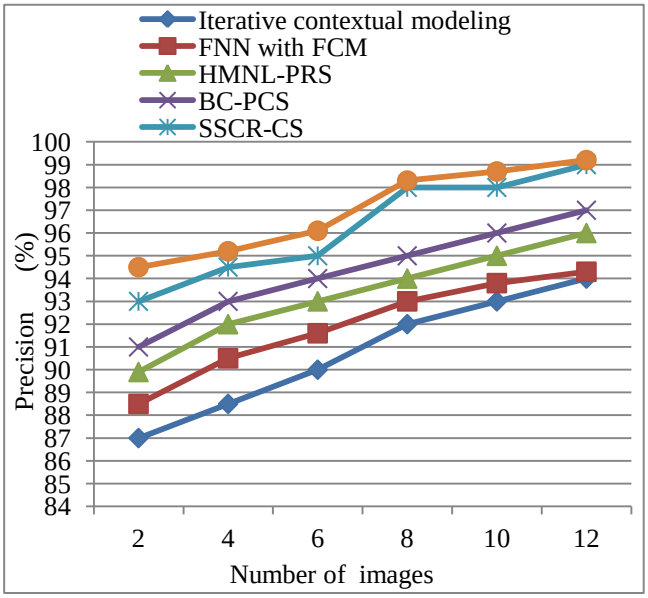

Figure 7. Precision comparison

The correlation of an iterative logical displaying, FNN with fluffy c-implies, HMNL-PRS, BC-PCS, SSCR-CS and HFE-MCSVM technique is exhibited in Figure 7 concerning accuracy. In X pivot the quantity of images is plotted and the exactness is plotted in Y hub. From this examination assessment it is demonstrated that the proposed HFE-MCSVM based character acknowledgment technique is exhibited the high exactness results.

C. Recall

The quantity of genuine positive archives recuperated over a hunt parceled by the absolute number of open related records. This procedure is known as Recall. The estimation of review worth is composed as pursues:

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive+False negative}}$$

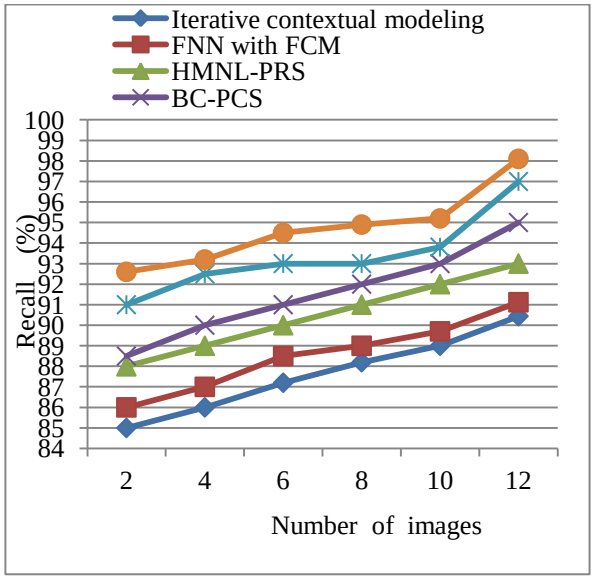

Figure 8. Recall comparison

The correlation of an iterative logical displaying, FNN with fluffy c-implies, HMNL-PRS, BC-PCS, SSCR-CS and HFE-MCSVM procedure is exhibited in Figure 8 regarding accuracy. In X pivot the quantity of images is plotted and the review is plotted in Y hub. From the assessment it is demonstrated that the HFE-MCSVM based character acknowledgment system has exhibited the high review results.

D. F-Measure

The joined estimation of the exactness and review as the consonant mean of accuracy and review is evaluated procedure is known as F-measure. The f-measure is composed as pursues:

$$F = 2.\frac{Precision.Recall}{Precision+Recall}$$

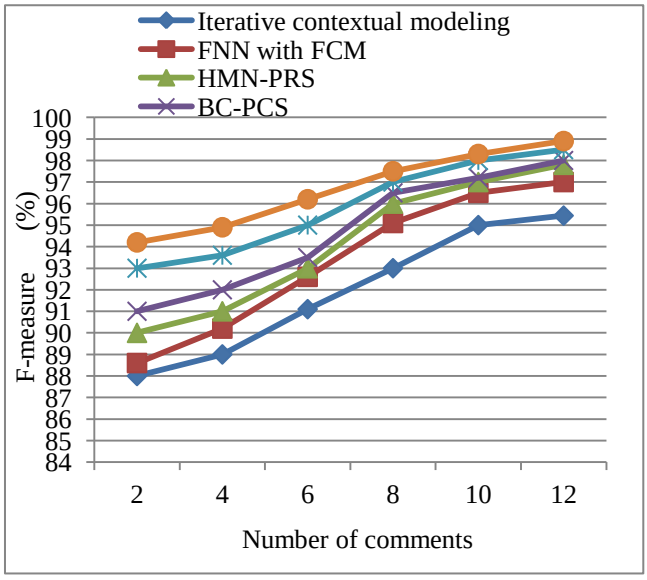

Figure 9. F-measure comparison

The consequence of the F-proportion of the new HFE-MCSVM based character acknowledgment system is contrasted with the other strategy like iterative relevant displaying, FNN with FCM, HMNL-PRS and BC-PCS based character acknowledgment philosophy. In Figure 9 illustrates the graphical portrayal of correlation of f-measure. The outcomes are shown that the new HFE-MCSVM based character acknowledgment system is high f-measure worth contrasted with the former character acknowledgment philosophy.

V. CONCLUSION

In this research work, initially image pre-processing is performed by using Gabor filter. After pre-processing segmentation of characters is performed by using overlapped character segmentation method. After segmentation character recognition is done by introducing the method namely hybrid feature extraction with Multiclass SVM classification approach. The overall assessment of the research work is done in the matlab simulation environment and then it is proved the proposed HFE-MCSVM shows better performance.

ACKNOWLEDGMENTS

This work is carried out with the financial support of University Grant Commission (UGC) under minor Research Project(MRP)

REFERENCE

1. Meng, L., Aravinda, C. V., Reddy, K. U. K., Izumi, T., & Yamazaki, K. (2018, October). Ancient Asian Character Recognition for Literature Preservation and Understanding. In Euro-Mediterranean Conference (pp. 741-751). Springer, Cham.
2. Stadermann, J., Jager, D., & Zernik, U. (2017). U.S. Patent Application No. 15/620,733.
3. Panagiotopoulos, P., Barnett, J., Bigdeli, A. Z., & Sams, S. (2016). Social media in emergency management: Twitter as a tool for communicating risks to the public. Technological Forecasting and Social Change, 111, 86-96.
4. Islam, N., Islam, Z., & Noor, N. (2017). A survey on optical character recognition system. arXiv preprint arXiv:1710.05703.
5. Al-Maadeed, S., Issawi, F., & Bouridan, A. (2017, May). Word Retrieval System for Ancient Arabic Manuscripts. In 2017 9th IEEE-GCC Conference and Exhibition (GCCCE) (pp. 1-5). IEEE.
6. Fung, P., Dey, A., Siddique, F. B., Lin, R., Yang, Y., Bertero, D., ... & Wu, C. S. (2016, December). Zara: A virtual interactive dialogue system incorporating emotion, sentiment and personality recognition. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations (pp. 278-281).
7. Rose, G. (2016). Rethinking the geographies of cultural 'objects' through digital technologies: Interface, network and friction. Progress in Human Geography, 40(3), 334-351.
8. Gupta, A., Vedaldi, A., & Zisserman, A. (2018). Learning to read by spelling: Towards unsupervised text recognition. arXiv preprint arXiv:1809.08675.
9. Shankar, S., Kumar, R. N., Mohankumar, P., Karthick, J., & Pradeep, S. (2018). Association of Job and Demographical Risk Factor with Occurrence of Neck Pain Among Hand Screen Printing Workers. In Ergonomics in Caring for People(pp. 131-137). Springer, Singapore.
10. Ogata, T., Torihata, T., & Takada, N. (2018). U.S. Patent No. 9,879,145. Washington, DC: U.S. Patent and Trademark Office.
11. Carston, R. (2016). Linguistic conventions and the role of pragmatics. Mind & Language, 31(5), 612-624.
12. Raghupathy, K. B., & Chandrasekaran, S. (2019). Benchmarking on offline Handwritten Tamil Character Recognition using convolutional neural networks. Journal of King Saud University-Computer and Information Sciences.
13. Panyam, N. S., Krishnan, R., & NV, K. R. (2016). Modeling of palm leaf character recognition system using transform based techniques. Pattern Recognition Letters, 84, 29-34.
14. Sadanand, A. K., Prashant, L. B., Ramesh, R. M., & Pravin, L. Y. (2015). Offline MODI character recognition using complex moments. Procedia Computer Science, 58, 516-523.
15. Bhunia, A. K., Roy, P. P., Mohta, A., & Pal, U. (2018). Cross-language framework for word recognition and spotting of Indic scripts. Pattern Recognition, 79, 12-31.
16. Lakshmi, T. V. (2018). Reduction of features to identify characters from degraded historical manuscripts. Alexandria engineering journal, 57(4), 2393-2399.
17. Premaratne, H. L., Järpe, E., & Bigun, J. (2006). Lexicon and hidden Markov model-based optimisation of the recognised Sinhala script. Pattern recognition letters, 27(6), 696-705.
18. Likforman-Sulem, L., Maître, H., & Sirat, C. (1991). An expert vision system for analysis of Hebrew characters and authentication of manuscripts. Pattern recognition, 24(2), 121-137.
19. Choudhary, P., Nain, N., & Ahmed, M. (2015). A structure for annotation and ground-truthing of Urdu handwritten text image corpus. Procedia-Social and Behavioral Sciences, 198, 84-88.

AUTHORS FROFILE

P. Balasubramanie is currently working as a Professor in the department of computer Science & Engineering, Kongu Engineering College, Perundurai, Tamil nadu, India. He is one of the approved supervisor of Anna University, Chennai and guided 26 PhD scholars. Currently he is guiding 9 PhD scholars. He has published 218 articles in International/National journals. He has authored/co-authored 11 books with the reputed publishers. Three of the books published are used as text/reference books by many of the leading Universities in India. He has completed one AICTE RPS as a Principal investigator and currently he is working as a principal investigator in a MRP of UGC. He has received Rs. 13 Lakhs of grant from various funding Agencies like AICTE, CSIR, NBHM, DRDO, INSA and so on and organized 21 STTP/SDP/Seminar/Workshops for the benefit of Faculty members and Research scholars. He has received 17 awards so far from various agencies. His area of interest includes Data mining, Networking, cloud computing and Optimization algorithms.

Dr. E.K.Vellingiriraj completed BSc (Computer Science) in Bharathiar University in 1999, then received his M.C.A. degree from Periyar University in 2002 and Completed M.E. Software Engineering from Anna University, Coimbatore, India in 2010. He completed MBA (HR) in Indira Gandhi National Open University (IGNOU), New Delhi in 2010. He is very much interested in Tamil language which led him to complete M.A. Tamil in TNOU, Chennai. He was completed his research in Natural Language Processing of Character Recognition from Historical Palm Manuscripts and Stone Inscriptions in Anna University, Chennai. Totally he has 13 years experience including 3 years industrial experience and 10 years of teaching. He has created his own Tamil BlogSpot and developed various Android apps for Tamil people and others.