International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

A Chronological Method of Detecting Image Based Email

Mallikka Rajalingam, M. Balamurugan

Abstract: ***In this paper we present a Visual feature extraction using improvised SVM and KNN classifiers. The proposed method is an automatic, stable, quick response automatic segmentation, followed by feature extraction and classification to detect spam from the images and the text. The KNN classifier is used to extract features by predicting nearest neighbour while SVM, analyze the data for classification and regression. The hybrid-based Visual feature extraction and classification is elaborated wherein this work discuss the proposed approach which incorporated using improvised SVM and KNN classifier. Moreover, identified patterns via feature extraction method by means of a minimum number of features that are effective in discriminating pattern classes. With all the aforementioned concepts elaborated, the experimental set-up was elaborated with the experimental task, and the results of the character recognition component are further elucidated.***

Index Terms: ***Detection, Image spam email, Recognition, and Segmentation.***

I. INTRODUCTION

The present advancement in internet, there is increased utilization of email communication which has become one among the fastest modes of communications. However, an increase in the usage of email communication has led to the increased rate of spam based issues all over the world. According to researchers [5] around 90 percent of emails that arrive at the mailbox of email users are spam emails wherein these emails contain junk information that tends to affect the normal computing utilities of email users. While spam emails are generally based on advertising content, in many cases they also contain malicious code and virus which might harm the users' account [2]. With advancements in the technologies to detect email spam emerged, spammers developed the concept of image spamming which tend to complicate the processes of detecting spam in image mails. Though previous researchers attempted to develop novel techniques for the detection of image spam, there is still a gap to develop an efficient image spam detection system which could detect spam in images wherein the scalability of the method should improve despite the type of image spam that is sent.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Mallikka Rajalingam*, Research Officer, Department of Computer Science & Engineering, Bharathidasan University, Trichy,(T.N.). India.
Dr. M. Balamurugan, Professor, Department of Computer Science and Engineering of Bharathidasan University, Trichy,(T.N.) India.

In this regard, the present work in this paper discussed about segmentation and recognition techniques involved in image spam detection. A hybrid character segmentation algorithm is proposed which combines several aspects such as DWT, Hough transform and pixel count analysis. Based on the examination of previous researchers, the aforementioned techniques were selected and were hypothesized to improve the segmentation efficiency. A sequential character recognition algorithm is projected which combines various techniques like contour analysis with improved local binary pattern proposed for text recognition and visual feature extraction. Based on the examination of literature article, the aforementioned techniques were selected and were hypothesized to improve the efficiency of feature extraction. In order to get the smooth contours of images, a double filter bank, the Laplacian Pyramid (LP), followed by Directional Filter Bank (DFB) provides better multi-scale decomposition and removes the low frequency. A spam image carries a message which is intended to reach client systems and displays the same [4]. Integrated approach [3] declared Image spam as a kind of email spam in which the textual message is implanted within an image submitting it as a picture. A study [7] recommended an image anti-spam system that utilizes diverse approaches of image element extraction and a fake neural model to categorize emails. After segmenting each character, the contour study is used to recognize a perfectly matched character. With the help of existing pattern stored in training phase, contour characters are matched by complex vector illustration [8]. Researchers [13] examine and compare two methods for image spam detection. First, they believe principal component analysis (PCA), resolve eigenvectors equivalent to a set of image spam and calculate scores by projecting images onto the resulting eigenspace. Next approach, on the extraction of image features and selection of a best possible split using support vector machines (SVM). Author [14] projected SVM and k-nearest neighbor (k-NN) based instance strength of intelligence approach. The approach tries to choose the examples that are close up to the option and that are efficiently labeled. The primary attention is to determine close up neighbors to a query test and organize a nearby SVM that conserve the partition done on the get-together of neighbors. For experimentation, Dredze dataset was used to show improvement in results. Performing image denoising [15] without using pre-arranged basic functions to signify the image, they used dictionary learning and sparse illustration.

The forthcoming segment of this paper is as follows: proposed work is briefed in section 2. Experimentation and performance analysis is explained in section 3. The presented model is concluded in section 4.

II. PROPOSED WORK

With the increasing significance of internet around the world, email is one of the foremost approaches of communication amongst people. But due to the flood of online data, most of the people`s inbox space is overwhelmed by unsolicited commercial email or spam email. The spam emails not only waste computing resources and network bandwidth of the internet users', but it also, affects the larger scale, interrupt enterprises' of standard system process. In order to detect image spam email, a novel framework is proposed which is combination of character segmentation, recognition and classification technique (CSRC). The proposed framework exploits to take an advantage of processing low level features and extraction of embedded text data. First, extracted the text character from image by segmentation process which includes combination of DWT and skew detection. Furthermore, applied logical AND as well as morphological dilation operators towards remove the non-text regions. The size of input image is reduced by applying a fusion of Hough transform along with spatial frequency cross correlation approach. Whereas the fusion based approach deliberates both texture and structure of input image. This segmentation process is utilized towards isolate the image features of a specific shape and detects the regular curves like circles, lines, ellipses and so on. Second, have recognized the character via Text recognition and Visual feature extraction approach which relies on contour analysis with improved LBP. This approach is robustness in contrast to illumination variations and its computational simplicity. Moreover, described the local spacial structure and determined the LBP of each pixel value. Third, the extracted text features are classified using KNN and SVM classifiers. Here, the KNN is utilized towards extract the text features through predicting nearest neighbour whereas SVM is used to analyze the text data for classification and regression process.

In our study, the experiment has been done based on the Dredze data set. This data set contains 3299 spam images and 2021 ham images. However, in processing stage we have eliminated the image which does not provide the enough information like image with no texture information and image size of less than 10 bytes. As a result, we have considered 2173 spam images and 1248 ham images for testing and validation of proposed CSRC framework. In addition, we have considered different format of input images such as .jpg, .png and .bmp.

The image is pre-processed via binarization and thinning operation. The feature extraction process of chain-coding is done with thinning image and stored in a file before classifier used. The text segmented into lines and then individual characters. The text is scanned and a line in the image file is extracted. The resultant line is used as input to the character segmentation process and segments one character at a time. The extracted characters are still need to be recognized and the resultant image is used as input to the character recognition process.

Further the recognized character is given as input to the classification approach. Thus the proposed method is an automatic, stable, quick response automatic segmentation, followed by feature extraction and classification to detect spam from the images and the text. Fig. 1 shows feature extraction and classifier techniques implies in proposed work.

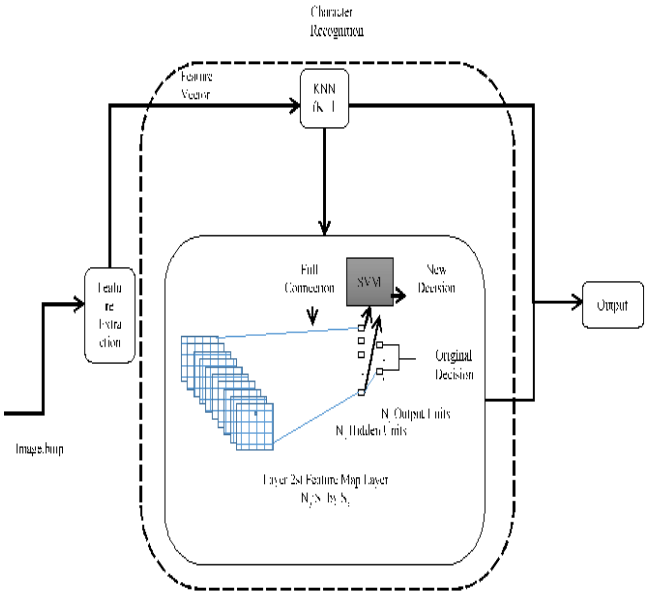

Figure 1. Feature extraction and classification techniques

A. *Support Vector Machine (SVM)*

Support vector machines employ a non-linear mapping to alter the initial training data into a greater extent. It seeks for the linear maximum dividing hyper-plane with a suitable number of support vectors (SVs). Info from two classes can forever be divided by a hyper-plane. The SVM discovers this hyper-plane employing support vectors and edges. Although the training time of even the quickest SVMs can be very high; they are extremely precise because of their skill to sample intricate non-linear decision borders. An SVM is fundamentally a binary classifier with the separate operation being the weighted mixture of kernel operations overall training models. The weights (coefficients) are studied by quadratic programming with the objective of magnifying the margin in feature space. Following the study, the models of non-zero weights are known as support vectors (SVs) that are saved and employed in categorization. Besides, we have employed linear kernel operation. Here, the kernels change the initial info into a greater extent feature space. Even if the initial info are non-linear, the changed info is divisible by a hyper-plane in feature space.

With this idea as the foundation, support vector machines have confirmed to attain fine generality functioning with no previous knowledge of the info. The rule of an SVM is to map the input stats onto a greater extent feature space non-linearly connected to the input space and establish a dividing hyper-plane with the supreme margin between the two classes in the feature space.

B. *Nearest Neighbor Search*

The K-Nearest Neighbor's algorithm is a procedure for categorizing objects grounded on the nearest training in the feature space.

Also, it's a kind of case-grounded studying by connecting unfamiliar design to the familiar based on certain distance.

To calculate a categorization operation by investigating the marked training points as nodes or anchor points in the n-dimensional space, where n is the feature dimension. Furthermore have estimated the Euclidean distance between the recollect point and pixel point then find k closest neighbors. Later, rank the attained distances in ascending order and take the mentioning points matching with the k smallest Euclidean distances. The KNN categorization separates information into a training data set and test data set.

III. EXPERIMENTATION AND PERFORMANCE ANALYSIS

For experimentation, 75 images were tested and were taken from the image spam dataset [1]. These images are in two groups that are 40 Spam and 35 Ham images. For the purpose of identifying both ham and spam images from the selected images of the Dredze dataset, both ham and spam images. In the present research, all images were taken for training and 35 images for testing. The training images comprise several images with different image characteristics. This is in line with the testing images wherein each image has its own characteristics when it comes to noise, background and so on. Furthermore, the use of 35 samples for testing is in line with the research which used only 20 images for both training and testing, however, revealed better results in terms of performance. All these images are selected from the dataset based on the various kinds of images present in the dataset, the uniqueness of each image and to show the efficiency of the proposed ham/spam detection method. The k-nearest neighbour classifier has been employed here to categorize amongst various classes or sets of characters where the classifier is primarily trained by employing certain training data sets. For this objective, certain images with identified character are selected for training, and their arrays are ascertained employing the above procedure of value assessment are explained. The K-NN classifier gives the optimal when the number of neurons at the hidden layer is 20.

A summary of classification accuracies obtained for three datasets are discussed. It shows that MLP yields the highest performance in most cases. It gives recognition rate is 96.7%. In MLP, a number of neurons at the hidden layer is chosen experimentally for every dataset. The selection of feature sets, feature optimization, post-processing and/or pre-processing can significantly contribute to the classification accuracy for all classifiers. Fig. 2 shows simulated results of feature extraction.

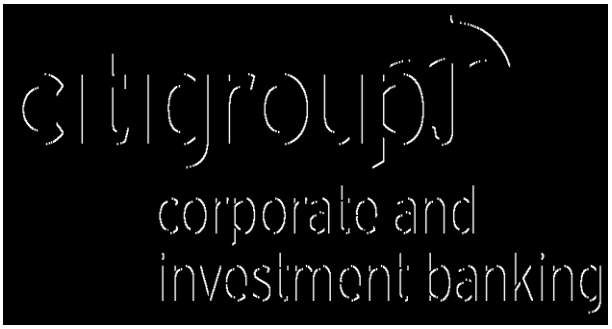

Figure 1. Simulated results of feature extraction

Figure 2. Simulated results of classifier technique

Fig. 3 shows simulated results of classifier technique. The text extraction and Character segmentation have been carried out by both groups like training and testing. The K-NN (K=1) with Euclidean distance functions was applied in the classification phase, and the multiclass SVMs was conducted. So as to confirm whether vectors relating to the upper and lower case variants of a similar letter is dispersed in neighboring locales of the element space or not by KNN classifier.

The more the two forms of the letter are comparable fit as a fiddle, the more their vectors are covering and ready to participate in a solitary class. The Average value of performance metrics obtained for this proposed algorithm is about 82.2 CR, 17.73 ER, 86.6% sensitivity, 81% specificity, 0.909 precision values, 0.866 recall, F-measure of about 0.883 and Accuracy of about 95.79%.The model improves output quality in terms of both sensitivity and specificity. Table I and table II shows results of spam and ham images.

Table I. Output results for SPAM images

SPAM							
Correct rate(CR)	**Error rate**	**Sensitivity**	**specificity**	**precision**	**Recall**	**F-measure**	**Accuracy**
82.3	17.7	100	78	0.909	1	0.95	96.7
85.4	14.5	100	82	0.909	1	0.9523	96.77
82.26	17.74	83.33	82	0.9091	0.833	0.87	95.16
77.42	22.6	75	78	0.91	0.75	0.822	96.8
83.9	16.13	75	86	0.91	0.75	0.823	93.56

Table II. Output results for HAM images

HAM							
Correct rate(CR)	**Error rate**	**Sensitivity**	**specificity**	**precision**	**Recall**	**F-measure**	**Accuracy**
82.261	17.72	100	79.1%	0.89	1	0.9524	95.161
82.321	17.62	100	78	0.902	1	0.958	95.6
74.2	25.81	100	68	0.91	1	0.95	95.2
82.25	17.75	66.6	86	0.88	0.67	0.75	95.13
80.64	19.35	91.67	78	0.863	0.916	0.9128	93.548

Similarly, the output of the five different HAM images after the segmentation process is tabulated in the table II which has the Average value obtained with these proposed algorithms are of 82.2 CR ,17.73 ER,86.6% sensitivity,81% specificity,0.909 precision, 0.866 recall, F-measure of about 0.883 and Accuracy of about 95.79%.

Table III. Comparison with existing approach

Author	Method	Accuracy
Zhang et al. (2014) [9]	ANN,SVM, decision tree	94.38,94.42 , 94.27
Wu (2009) [10]	ADTree	91.60
Lekha and Prakasam (2016) [11]	SMO	92.63
Sharma and Arora (2013) [12]	RANDOMTREE	91.54
Rusland et al. (2017) [6]	Naıve Bayes classifier	82.88%
Author (2019)	Proposed method	95.798

An algorithm which has a highest accuracy will be considered as the enhanced approach with better classification capability. Therefore, table III compared the accuracy range of few existing classifier studies. Though ANN, SVM, decision classifier method developed [9] has maximum accuracy 94% than other classifiers, it consumes more time to build model. Moreover, Naive Bayes framed [6] least accuracy of about 82.88% than other classifiers. Hence, our proposed method achieves the accuracy rate of about 95.79% and consumes less time from which it is concluded that it is best classifier concerning accuracy. The pictorial representation of the performance comparison of the proposed method with the conventional classifier methods. The conventional techniques such as the Naïve Bayes, AD tree, SMO, Random Tree, ANN, SVM, Decision tree which has the accuracy rate of about 82.88%, 91.60%, 92.63%, 91.54%, 94.38%, 94.42%, 94.27%. However, our proposed technique achieved the enhanced accuracy rate of about 95.80% which represents that it outperformed the conventional methods. Fig. 4 illustrations proposed method compared with other existing method.

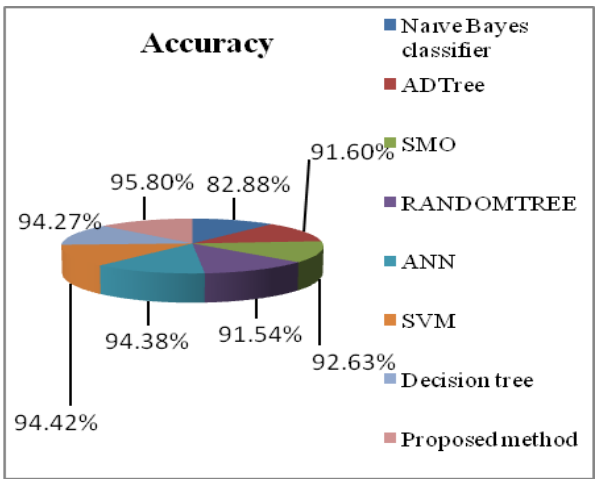

Figure 4. Performance comparison with existing method

IV. CONCLUSION

This paper explains the results of segmentation, recognition and classification techniques are applied for character detection to identify the SPAM email.

The accuracy of the classification techniques are evaluated by receiver operating characteristics which includes the measures of precision, recall, f-measure, CR, ER, sensitivity and specificity. Finally provides a comparative study of how all the techniques discussed so far are used for improving classification accuracy. To combine the text and document reconstruction approach with character classifier towards attains a weighted decision about the class. Also extract the non-text-content extracted features like hyperlink, header and embedded image data will be the future enhancement of the work.

REFERENCES

1. Dredze, M., Gevaryahu, R. & Elias-Bachrach, A. (2007). Learning Fast Classifiers for Image Spa. In: *proceedings of the Conference on Email and Anti-Spam*. 2007, CEAS.
2. Firte, L., Lemnaru, C. & Potolea, R. (2010). Spam detection filter using KNN algorithm and resampling. In: *Proceedings of the 2010 IEEE 6th International Conference on Intelligent Computer Communication and Processing*. [Online]. August 2010, IEEE. Available from: http://ieeexplore.ieee.org/document/5606466/.
3. Kumar, P. & Biswas, M. (2017). SVM with Gaussian kernel-based image spam detection on textual features. In: *2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT)*. [Online]. February 2017, IEEE, pp. 1–6. Available from: http://ieeexplore.ieee.org/document/7977283/.
4. Mehta, B., Nangia, S., Gupta, M. & Nejdl, W. (2008). Detecting image spam using visual features and near duplicate detection. In: *Proceeding of the 17th international conference on World Wide Web - WWW '08*. [Online]. 2008, New York, New York, USA: ACM Press, pp. 497–506. Available from: http://portal.acm.org/citation.cfm?doid=1367497.1367565
5. Rekha & Negi, S. (2015). A Review on Different Glaucoma Detection. *International Journal of Engineering Trends and Technology*. 11 (6). pp. 2–7.
6. Rusland, N.F., Wahid, N., Kasim, S. & Hafit, H. (2017). Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets. *IOP Conference Series: Materials Science and Engineering*. [Online]. 226. p.p. 12091. Available from: http://stacks.iop.org/1757-899X/226/i=1/a=012091?key=crossref.262a040d608eaab52b8501086da85f26.
7. Sanches, B.C. & Moreira, E.M. (2017). *Detecting image spam with an artificial neural model*. 15 (1). pp. 296–315.
8. Soumya, K.R., Babu, A. & Therattil, L. (2014). License Plate Detection and Character Recognition Using Contour Analysis. *International Journal of Advanced Trends in Computer Science and Engineering*. 3 (1). pp. 15–18.
9. Zhang, Y., Wang, S., Phillips, P. & Ji, G. (2014a). Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. *Knowledge-Based Systems*. [Online]. 64. pp. 22–31. Available from: http://dx.doi.org/10.1016/j.knosys.2014.03.015.
10. Wu, C.H. (2009). Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. *Expert Systems with Applications*. [Online]. 36 (3 PART 1). pp. 4321–4330. Available from: http://dx.doi.org/10.1016/j.eswa.2008.03.002.
11. Lekha, K.C. & Prakasam, S. (2016). Prediction of Respondants' Knowledge towards Cyber Security measures using various Classification Algorithms. *International Journal of IT and Knowledge Management*. [Online]. 10 (1). pp. 1–5. Available from: http://csjournals.com/IJITKM/PDF10-1/1. Chitra.pdf.
12. Sharma, S. & Arora, A. (2013). Adaptive Approach for Spam Detection. *IJCSI International Journal of Computer Science*. [Online]. 10 (4). Available from: https://pdfs.semanticscholar.org/956c/dfa8574d01f0cdb2eaa5383ea5028a1eadc6.pdf.
13. Annadatha, A.S. (2018). Image spam analysis and detection. *Journal of Computer Virology and Hacking Techniques*. 14(1). pp. 39-52. Available from: https://link.springer.com/article/10.1007/s11416-016-0287-x.
14. Zamil, Y.K., Suhad A. Ali, & Abdullah Naser, M. (2019). Spam image email filtering using K-NN and SVM. *International Journal of Electrical and Computer Engineering (IJECE)*. 9(1). Pp. 245-254.
15. Bhagya Prasad Bugge, Bh.S.S.D.S Nagendra Varma, & A.Amarnath. (2019). Image Denoising and Metric Parameters Improvement using Dictionary Learning and Sparse Coding. *International Journal of Recent Technology and Engineering (IJRTE).* 8(153). pp. 1-5.

AUTHORS PROFILE

Mallikka Rajalingam received her M.Sc Information Technology from Bharathidasan University, Tiruchirappalli, India in 2005, M.Phil Computer Science from Madurai Kamaraj University, Madurai, India in 2008, M.Tech Computer Science & Engineering from SASTRA University, Thanjavur, India in 2009. She worked as a Research Officer (RO) at School of Computer Science, Universiti Sains Malaysia (USM), Malaysia. She is currently pursuing the Ph.D. degree at the Department of Computer Science & Engineering, Bharathidasan University, Trichy, India. Her research interests include image processing, computer vision, pattern recognition, character recognition, document image analysis, text analysis and multimedia networking.

Dr. M. Balamurugan is currently working as Professor and Head in the Department of Computer Science and Engineering of Bharathidasan University, Trichy, India. He has credits of 20+ international and national conferences publications. He has published 30+ research papers in national and international journals. His research interests are mainly focused on the area of Data Science. He has supervised several research scholars in these areas.