2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies

Optical Character Recognition for English and Tamil Using Support Vector Machines

R.Ramanathan, S.Ponmathavan, N.Valliappan
L.Thaneshwaran, Arun.S.Nair
Department of Electronics & Communication Engineering,
Amrita Vishwa Vidyapeetham,
Coimbatore, India
e-mail: r_ramanathan@ettimadai.amrita.edu

Dr. K.P.Soman
Center for Excellence in Computational Engineering and Networking (CEN)
Amrita Vishwa Vidyapeetham,
Coimbatore, India
Email: kp_soman@amrita.edu

Abstract: **Optical Character Recognition is an evergreen area of research and is verily used in various real time applications. This paper proposes a new technique of Optical character Recognition using Gabor filters and Support Vector machines (SVM). This method proves to be very effective with the use of Gabor filters for feature extraction and SVM for developing the model. The model proposed is trained and validated for two languages – English and Tamil and the results are found to be very much encouraging. The model developed works for the entire character set in both the languages including symbols and numerals. In addition , the model can recognise the characetrs of six different fonts in English and Twelve different fonts in Tamil. The average accuracy of recognition for English is 97% and for Tamil it is 84%, which is achieved in just three iterations of training. The method can turn out to be a suitable candidate for future applications in this area.**

Keywords: Gabor Filters, Optical Character Recognition, Support Vector Machines

I. INTRODUCTION

Character recognition has a great potential in data and word processing for instance, automated postal address and ZIP code reading, data acquisition in bank checks, processing of archived institutional records, etc. In the Recent years, optical character recognition (OCR) has gained a momentum since the need for converting the scanned images into computer recognizable formats such as text documents has increased applications. OCR is one of the most fascinating and challenging areas of pattern recognition with various practical applications.

The process of character recognition involves extraction of defined characteristics called features to classify an unknown character into one of the known classes. Therefore, OCR involves two processes: 1.Feature extraction 2.Classification. The process of character recognition becomes very tough in the case of Indian languages like Tamil. Many inter-class dependencies exist in Tamil. In Tamil language, many letters look alike. So classification becomes a big challenge. [10,11]also discusses about the Character Recognition methods. In image processing, gabor filters are extensively used for feature extraction [1, 2]. Also in recent years, Support Vector Machines has gained momentum in the field of classification. SVM has good generalization capabilities, which are very essential in OCR [8]. In addition, the combination of SVM and gabor filters has given good results for font recognition [5, 6] and facial recognition [1, 3]. In this paper, a method is proposed for character recognition, which uses Gabor filters to extract features and SVM for classification. The proposed method validates both training and testing methodologies. In training phase, the features extracted using Gabor filter is fed to SVM along with respective class values. This will train the SVM model. Whereas in testing phase the features are extracted by Gabor filter and SVM tests those features and assigns respective class values These class values can be directly converted into the respective character or its ASCII value in a text file. In addition, the discussed methodology is verified both for English and Tamil languages.

II. INTRODUCTION TO GABOR FILTER

Gabor filters possess optimal localization properties in both spatial and frequency domain. Gabor filters give a chance for multi-resolution analyses by giving coefficient matrices [6]. Here a 2D Gabor filter has been used for the purpose of feature extraction. A 2D Gabor is a Gaussian modulated sinusoid in the spatial domain and a shifted as a shifted Gaussian in the frequency domain. It is represented by:

$$g_{\gamma,\eta,\varphi,\lambda} = \exp(\frac{x'^2 + \gamma 2 y'^2}{2\sigma^2}).\cos(\frac{2\pi x'}{\lambda} + \varphi) \qquad (1)$$

$$x' = x\cos\theta - y\sin\theta$$
$$y' = x\sin\theta - y\cos\theta \qquad (2)$$

The Gabor filters can be better used by varying the parameters like λ, γ, φ and θ. In the above equations, x and y represent image coordinates; s is the standard deviation of Gaussian function which is usually set to 0.56 λ; λ is the wave length of cosine equation; γ characterizes the shape of Gaussian, circular shape for γ=1 and elliptic for γ<1 and θ represents the channel orientation and takes values in interval (0, 360). Since it is symmetric, θ varies from zero to 180.

978-0-7695-3915-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ACT.2009.155

III. INTRODUCTION TO SVM

SVM is the method of creating functions from a set of labeled training data [4]. The function can be either a classification function or a general regression function. SVM finds an optimal separating hyper-plane between data points of different classes in a high dimensional space. Finding an optimal hyper-plane for non-linear patterns is the solution of the following optimization problem.

$$Min..\frac{1}{2}w^T w+\frac{C}{2}\varepsilon^T \varepsilon \tag{3}$$

Such that

$$D(w^T \varphi(x)-\gamma e)+\varepsilon \le e \tag{4}$$

where W is the coefficient vector, ϕ(x) is the non-linear mapping function that maps input vector to higher dimensional space, γ is the bias term, C is the cost factor, ξ is the slack variable, D is the diagonal matrix containing class values.

IV. TRAINING SVM MODEL

In this paper the character recognition model for two languages, English and Tamil are discussed. For English, 94 symbols including alphabets (small, caps), numbers and symbols of 10 different fonts of font size 14 were taken. In addition, the all font styles like bold, italic, bold italic, regular is used. This will train the model for all conditions. For Tamil, 156 symbols including alphabets, numbers of 12 different fonts of font size 14 were considered and trained separately. The training of the model involves the following steps. Thus, the model developed will be robust and work for the standard fonts in English and Tamil that are frequently used in Texts. Figure 1 gives the steps involved in training.

A. Scaning

The training files are made ready for all the fonts of various styles [10 images for English, 12 images for Tamil]. Then the training files are scanned and saved as bitmap images.

B. Preprocessing

The scanned image may have some skewness and the whole image on 3-d scale cannot be processed. Therefore, the image is normalized and skewness is corrected and is converted to a binary images.

C. Segmentation

The first process on preprocessed image is segmentation where we use the following algorithm to extract all characters from the image. Segmentation was performed in two phases. (i) Line segmentation wherein each line in the document was segmented using horizontal profile. (ii) Character segmentation wherein each character in the line was segmented using 8-connected component analysis [Haralick *et al.,* 1992]. The two-phase approach was adopted based on a comparative study where this approach yielded a better result.

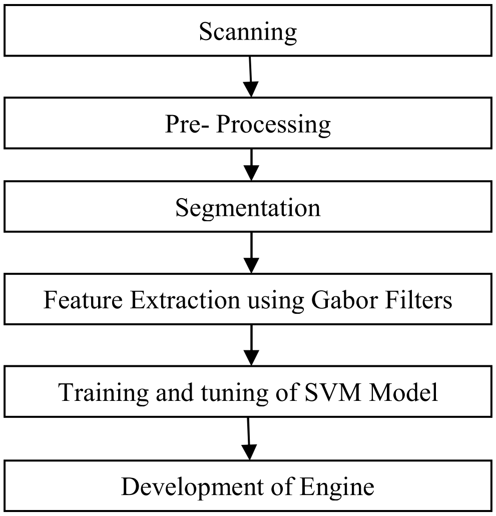
Scanning
Pre- Processing
Segmentation
Feature Extraction using Gabor Filters
Training and tuning of SVM Model
Development of Engine

Figure 1. Flow chart for training

D. Feature Extraction Using Gabor

After the segmentation process, the segmented images are resized into 64x64 binary images. These images are passed through a bank of 24 Gabor filters. The λ values of (2.7, 4.1, and 5.4) are fixed. The θ is taken as kπ/8, where k is varied as 1, 2…8.Thus 24 Gabor channels are obtained. Figure gives the bank of Gabor filters. For each image, 50 features are obtained by passing through these filters. Thus, a feature vector containing 50 features are obtained in this step.

E. Training SVM Model

Each feature vector is appended with its class value. These vectors are given to the Support Vector Machine so that it creates a decision boundary that has the maximum distance to the closest points in the data set. In this process data set are mapped to higher dimensional space by using a RBF kernel. C-SVM has been used in this paper

F. Tuning Of Parameters

In the above classification using C-SVM, the parameters like cost factor (c) and gamma (g) of the kernel function should be tuned for getting maximum accuracy. Any special optimization algorithm can be employed to find the best value for the parameters.

G. Development Of Engine

Once the tuning process is over, the SVM model is ready to classify any kind of new input. The engine can be developed by implementing the SVM model using any programming language.

V. TESTING THE MODEL

A newly scanned image will act as the test image. The test image undergoes the following process:

1. Scanning
2. Pre-processing

3. Segmentation
4. Feature Extraction using gabor
5. Classify using SVM Model

The first four steps are same as discussed earlier. The only change is in the last step. Once the feature vector is obtained, it is passed through the SVM model. The output of this stage gives the class values. This class values can be mapped to corresponding alphabet.

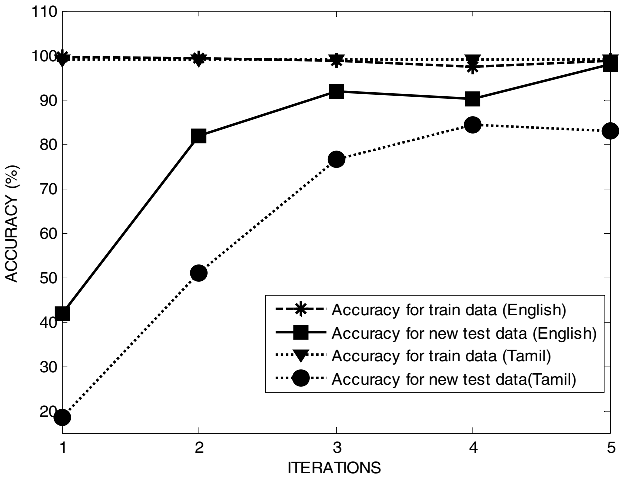

Figure 2. Accuracy Plot of the Proposed System

TABLE I. ACCURACY FOR ENGLISH AND TAMIL

	Training data	Accuracy for training data	Accuracy for new testing data
English	Iteration 1	99.47%	41.73%
	Iteration 2	99.17%	81.92%
	Iteration 3	98.85%	91.94%
	Iteration 4	97.25%	90.26%
	Iteration 5	98.83%	97.78%
Tamil	Iteration 1	98.95%	18.42%
	Iteration 2	98.98%	51.09%
	Iteration 3	99.05%	76.64%
	Iteration 4	99.05%	84.31%
	Iteration 5	99.04%	82.98%

VI. RESULTS AND DISCUSSION

The proposed model is trained for both Tamil and English separately. The model is checked by giving test images of different fonts. The table I above gives the accuracy for English. The accuracy plot against the number of iterations is also depicted in figure 2.It can be inferred that the English character recognition model works considerably well with an average accuracy of 97%. It can be seen that four iterations are used to reach the accuracy. As for the Tamil, the average accuracy is around 84%. However, a fact to be noted is the entire character set of Tamil language for 12 fonts have been considered and the said accuracy is achieved. The English character works with nearly six standard fonts incorporating the entire character set with symbols. It is noteworthy to obtain such a good accuracy in just three to four iterations of training.

VII. CONCLUSION

This paper presented an efficient algorithm for classification of characters using Gabor filters and SVM. The system was applied for the recognition of printed English and Tamil language characters, which included the entire character, set with symbols and numerals. The method works brilliantly for frequently used six English fonts and standard twelve different Tamil fonts. As shown in figure 4 , the accuracy increases with number of iterations and in the proposed method, a good accuracy is reached with few iterations The experimental procedures are explained and the result listed out depicts the efficiency of the system. The algorithm did prove more efficient and can be a suitable alternative for the Optical Character Recognition when compared to existing systems. Further optimization in the algorithm and development of engine is under progress.

REFERENCES

[1] Liu. C & Wechsler. H (2002). Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition. IEEE Trans. Image Processing, Vol. 11.467-476

[2] Praseeda Lekshmi V, Dr. M Sasikumar, Divya S Vidyadharan (2009). Facial Expression Classification from Gabor Features Using SVM. Computer Society of India journal for February 2009.

[3] R.Ramanathan, Arun.S.Nair, V.Vidhyasagar, N.Sriram, K.P.Soman "A Support Vector Machines Approach for Efficient Facial Expression Recognition" to be published in the Proceedings of International Conference on Advances in Recent Technologies in Communicaion and Computing – ARTCom 2009, India, 27-28 Oct 2009

[4] K.P.Soman, Loganathan.R, V.Ajay "Machine Learning with SVM and other Kernel Methods", Prentice Hall of India

[5] Borji, and M. Hamidi. "Support Vector Machine for Persian Font Recognition" Proceedings of world academy of science, Engineering and technology Volume 22,July 2007,ISSN1307-6884.

[6] R.Ramanathan, L.Thaneshwaran, V.Viknesh, T.Arunkumar, P.Yuvaraj , K.P.Soman "A Novel Technique for English Font Recognition Using Support Vector Machines" to be published in the Proceedings of International Conference on Advances in Recent Technologies in Communicaion and Computing – ARTCom 2009, India, 27-28 Oct 2009

[7] C.-C. Chang And C.-J. Lin, "Libsvm - A Library For Support Vector Machines," 2.82 Ed, 2001. Software Available At Http://Www.Csie.Ntu.Edu.Tw/~Cjlin/Libsvm.

[8] R.Ramanathan, N.Valliappan, S. Pon Mathavan, M.Gayathri, R.Priya, K.P.Soman "Generalized and Channel Independent SVM based Robust Decoders for Wireless Applications" to be published in the Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing – ARTCom 2009, India, 27-28 Oct 2009

[9] R.Ramanathan, P.A.Rohini, G.Dharshana., K.P.Soman "Investigation and Development of Methods to Solve Multi-Class Classification Problems " to be published in the Proceedings of International Conference on Advances in Recent Technologies in Communicaion and Computing – ARTCom 2009, India, 27-28 Oct 2009

[10] Qing Chen, Evaluation of OCR Algorithms for Images with Different Spatial Resolutions and Noises, Master thesis, School of Information Technology and Engineering, University of Ottawa, 2003

[11] MK Hu, Visual Pattern Recognition by Moment Invariants, *IRE Trans Information Theory*, vol. 8, pp. 179-187, 1962