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

Efficient Technique for word identification and recognition in Telugu Documents

Kesana Mohana Lakshmi, Tummala Ranga Babu

Abstract*: *Telugu language is one of the most spoken Indian languages throughout the world. Since it has an old heritage, so Telugu literature and newspaper publications can be scanned to identify individual words. Identification of Telugu word images poses serious problems owing to its complex structure and larger set of individual characters. This paper aims to develop a novel methodology to achieve the same using SIFT (Scale Invariant Feature Transform) features of telugu words and classifying these features using BoVW (bag of visual words). The features are clustered to create a dictionary using k-means clustering. These words are used to create a visual codebook of the word images and the classification is achieved through SVM (Support Vector Machine).

***Index Terms*: Telugu, SHIFT, SVM, BOVW**

I. INTRODUCTION

Large amount of research and development in the area of machine learning and image processing these days. [1]Robust feature descriptors like HOG (Dalal and Triggs, 2005), [2]SIFT (Lowe, 2004) and GLOH have been developed and used in different applications like image recognition and image registration. Large number of digital libraries such as Universal Library (UL) [1],Digital Library of India (DLI) [2], and Google books are emerging for archival of multimedia documents. These documents cannot be stored as text always. This makes the search for relevant documents even more challenging. Nowadays, the storage devices are becoming cheaper and imaging devices are becoming increasingly popular. This motivates researchers to put efforts on developing efficient techniques to digitize and archive large quantity of multimedia data. The multimedia data includes text, audio, image and video. At this stage, most of the archived materials are printed books, and digital libraries are collection of document images. To be more precise, digitized content is stored as images corresponding to pages in books. These documents are typically available in very large numbers hence manually grouping and filing these documents for making them available easily is very tedious task. However, it is very important that these documents are made accessible to the users who would in fact like to search them with relative ease.

Fig 1: Composite Characters in Telugu

However, it is very important that these documents are made accessible to the users who would in fact like to search them with relative ease. Scanned or digital form of documents do not contain searchable text as it is but contain words as images which cannot be searched or retrieved by existing search engines. Traditional text search is based on matching or comparison of textual description (sayin ASCII/ UNICODE) in association with a powerful language model. These techniques cannot be used to access content at the image level, where text is represented as pixels but not as text. The best way of processing these digital documents is to segment the textual content present in the documents. Once the contents are separated out, a representational scheme (profile feature) can be applied to them to get their representational form, which could be ready to use in a content-based image retrieval system. Additionally, a direct method to access these documents is by converting document images to their textual form by recognizing text from images. The paper has been divided into four more sections. The next section lays out the work carried out on BOW and its theoretical framework. The third section discusses the methodology used in the work. The fourth chapter presents the results and discussion about it. The fifth section derives the conclusion from the work.

BoVW: Sivic introduced Bags-of-visual words and applied it in the field of video retrieval system. Mostly because of its efficiency and robustness, it was applied in other fields such as image retrieval and image categorization. The main aim in classification is to build a system that can basically assign objects to a certain category based on the input samples which have been given to it during the training phase. This can be easily done by representing the image as a combination of basic features (words) taken from a dictionary. The word used in case of visual classification is the image feature. Various image features can be used to create these words such as image patches, histograms, HOG, SIFT, GLAC, Gabor, log-Gabor etc.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Kesana Mohana Lakshmi*, Department of Electronics and Communication Engineering, CMR Technical Campus, Hyderabad, Telangana, India
Tummala Ranga Babu, Department of ECE, RVR&JC College of Engineering, Guntur, A.P, India

But mostly these features are not robust to noise and hence they are not directly used as words but assigned to a high dimensional space. The assignment of these words can be achieved by employing a vector quantization method like k-means and it is followed by a classifier which directly classifies images based on its constituent words. A most commonly used and successful classifier in the case of BOW is SVM. The output of this quantization process gives the dictionary. This concept takes inspiration from Word-wise Video Script Identification [11] where a document is described in term of words. Similarly, in this model, comprised of bag of words to analyze the information inside an image. A 'visual word' is represented by group of features that relate to properties of certain information identified as key points. These features are separated into classes. A 'visual word' is a vector that denotes the vector which gives the features of each class centroid and the group of classes are termed as the codebook. Particularly, every local point gives a visual word that relates to the nearest centroid determined by Euclidean distance. BoVW model [11] has been used in applications like classification of scenery images [22]. Rothacker [23] applied bag-of-features representation for the design of a semi-continuous probabilistic model for Arabic handwriting identification [8].

Rest of the paper includes the related work done in the field of word script identification and recognition, proposed methodology, simulation results followed by conclusion and references considered for the development of proposed methodology for word script identification with more accurate recognition.

II. RELATED WORK

Optical character recognizers (OCRs)are required to obtain the textual content from these documents [3, 4]. Success of OCR based text image retrieval schemes mainly depend on the performance of optical character recognition systems(OCRs) [5].In literature, the use and application of OCR systems are well demonstrated for many languages in the world [6]. For Latin scripts and some of the Oriental languages, high accurate recognition systems are commercially available for use. However, for Indian language with their own scripts, much attention has not been given for developing robust OCRs that successfully recognize diverse printed text images. Therefore, many alternate approaches are presented by researchers to access content of digital libraries in these languages [7]. The focus is on recognition free approaches for retrieval of relevant documents from large collections of document images. In recent years, retrieval of document images from the information of query word has emerged as a vital research field [8-12]. This is also a successful alternative development for the applications of recognition-based handwritten and printed documents in most of the complex scripts retrieval system. Technically, these approaches retrieve the relevant word images from the database by utilizing the extracted features of query word, which will be extracted based on various methodologies with similarity measurement. Based on all the survey papers discussed above and in the literature exposed that the unique modeling of structure of script is a challenging task for recognizing or identifying the word-script. Therefore, the present study investigated that a novel Telugu word recognition approach based on Bag-of-visual-Words (BovW) and K-means clustering with SIFT descriptor and SVM classification. We also presented a comparative study and analysis with the conventional techniques to understand the effectiveness and robustness of the proposed approach.

III. METHODOLOGY

In our proposed system, we have used patch-based SIFT descriptors along with a spatial pyramid matching approach for extraction of features from the segmented words. Figure 2 depicts the block diagram of the proposed method.

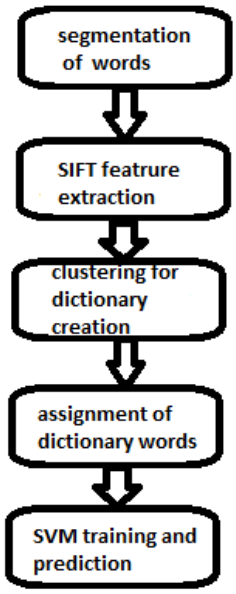


Fig 2: Block diagram of proposed method

Preprocessing: The major hurdle in the classification of Telugu documents is in tackling the skew introduced while converting the paper document to electronic format. In this paper, we have first preprocessed the image to de-skew the document. There are different ways described in the literature to de-skew documents. In this paper, we have used profile projection based de-skewing process. In this method, the input image is rotated across a number of angles and a Projection profile is determined at each angle. At each projection profile, features are extracted to calculate the skew angle. The baseline of the script can be identified as the region of the image having a maximum number of the black pixels.

Segmentation: The segmentation of words is done on the basis of region growing method. The image is first binarized and then morphological operators are applied to remove any noise in the image. The connected components are extracted and the noisy pixels are removed depending on the area of the text patches. The areas to be considered are at a minimum of twelve thousand pixels Bag of Visual Words.

Patch-Based SIFT Descriptor: A group of local image patches is taken using SIFT key point detector and a vector of visual descriptors is applied on every one of these patches individually.

SIFT feature extraction is done in two stages – detection and descriptor. In order to detect the keypoints, the scale-space extrema is determined. The scale space is a continuous function of scale σ . The maxima of LoG (Laplacian of Gaussian) give the best detail of the scale. The Gaussian kernel is given by

$$g_\sigma(x) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{1}{2}\frac{x^T x}{\sigma^2}\right). \quad (1)$$

Fig 3: Block Diagram of proposed method

The Gaussian scale space is the set of images smoothed using this Gaussian kernel

$$I_\sigma = g_\sigma * I, \sigma \geq 0 \quad (2)$$

In practice, a nominal value of σ is chosen to calculate the scale space

$$I_\sigma = g_{\sqrt{\sigma^2-\sigma_n^2}} * I_{\sigma_n}, \qquad \sigma \geq \sigma_n \quad (3)$$

Scales are sampled at logarithmic steps

$$\sigma = \sigma_0 2^{o+\frac{s}{S}}, \qquad s = 0, \dots, S-1,$$
$$o = o_{min}, \dots o_{min} + O - 1 \quad (4)$$

Where σ_0 is the base scale, o_{min} is the zeroth octave, O is the total number of octaves and S is the total number of scales in each octave. Key points are then identified as local extrema of the Difference of Gaussians (DoG) scale space, calculated by determining the difference of successive scales of the Gaussian scale space:

$$DOG_{\sigma(o,s)} = I_{\sigma(o,s+1)} - I_{\sigma(o,s)} \quad (5)$$

Out of these detected keypoints, low-contrast responses and edge points are removed. After the detection stage is completed, the descriptors are calculated for them. For this first, the image gradient vector is calculated first by

$$J(x,y) = \nabla I_\sigma(x,y) = \left[\frac{\partial I_\sigma}{\partial x}\frac{\partial I_\sigma}{\partial y}\right] \quad (6)$$

The descriptor is a 3D spatial histogram capturing the distribution of J(x,y).

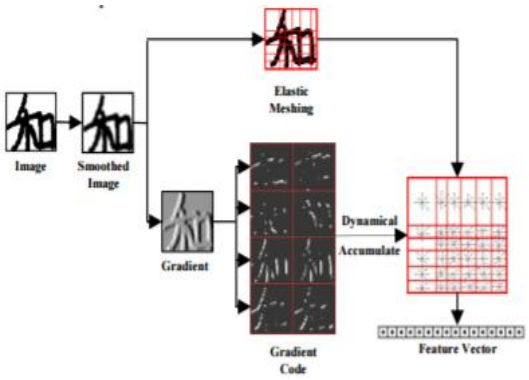


Fig 4: Calculation of SIFT Vectors

The descriptors are calculated using 4x4 blocks of cells, with the cells being 3x3 pixels at its finest scale. SIFT descriptors[20] have high invariance to translation and scale. Since it uses oriented gradients it also has invariance to intensity. Increasing the size of patch demonstrated poor feature extraction. The same scenario occurs when the grid spacing between the patches is increased too much. After the SIFT keypoints have been determined, these points are used to build the dictionary.

Codebook Generation: K-means clustering is applied on the SIFT keypoints to create the codebook. Since the SIFT features have been calculated over patches, the feature dimensions are too high. By using the clustering process, centers are calculated and for a cluster and the feature descriptors get converted into clusters. K random vectors are selected and the squared Euclidean distance of all the features from these K vectors are computed. Based on this process we get k clusters. A SIFT[20] descriptor vector X_j is put in the i^{th} cluster if the squared Euclidean distance of this vector from the randomly selected vector is minimum in that cluster. At every iteration, the centroid of these clusters is calculated. When the centroid distances satisfy the threshold value, the clustering process terminates. In k-means clustering, each codeword is represented by the cluster center which is the average of all features that belongs to this codeword. As a result, it is fast and simple to compute. Its main objective is to minimize the objective function also known as squared error function given by

$$J(V) = \sum_{i=1}^{c}\sum_{j=1}^{c_i}(\|x_i - v_j\|)^2 \quad (7)$$

Where $\|x_i - v_j\|$ is the Euclidean distance between x_i and v_j, c_i is the number of data points in the i^{th} cluster
c is the number of clusters

Dictionary Construction: BoVW model is a dictionary-based method which was first used to represent documents by considering each document as a "bag", which consists of many words from the dictionary (codebook). By using the similar idea for image representation, BoW has been utilized in computer vision field [14]quite a lot, especially for object categorization. Therefore, the image can be considered as a document, and feature-set determined from the image are "words". To represent the image as the BoVW model, a widely-used method is to extract the SIFT [20]as the image detector/descriptor. After extracting interesting points (features) from the training images, we convert feature vectors into "words" (code words). The most important factor while creating a dictionary is the size of codeblocks which will be used to represent the image contents. Based on experimental results, we can say that the larger the codeblock size the higher classification accuracy can be achieved. Selecting the appropriate size presents the tradeoff between discriminativity and generalizability. Using a reduced size, the visual-word is not very differentiable as different keypoints can correspond to the same visual word. As we increase the size of the vocabulary, the feature becomes quite differentiable but possesses lesser generalizability and less robust against noise because same keypoints will now point to different words. Using an enhanced vocabulary size also raises the computational cost of running classifiers. The goal of the dictionary construction is to identify a set of visual patterns which depict the image content. We have illustrated a dictionary extracted from the training set of 15000 Telugu words. These set of patterns are termed as visual words.

SVM Classifier: Since SVM was actually developed [6] for binary classification, in our work some multi-class SVM methods have been applied to realize multi-class classification. For L classes, this method uses L*(L-1)/2 SVM classifiers, where each of them is trained on data from two classes. The SVM classifier basically works on the concept of kernel functions. These functions describe the similarity relationship between the Telugu words to be classified. The image representation which we have calculated here is a spatial pyramid with term frequencies. A Radial Basis kernel function has been used to classify the word images.

$$K(H, H') = \exp\left(-\gamma D_{\cap}(H, H')\right) \quad (8)$$

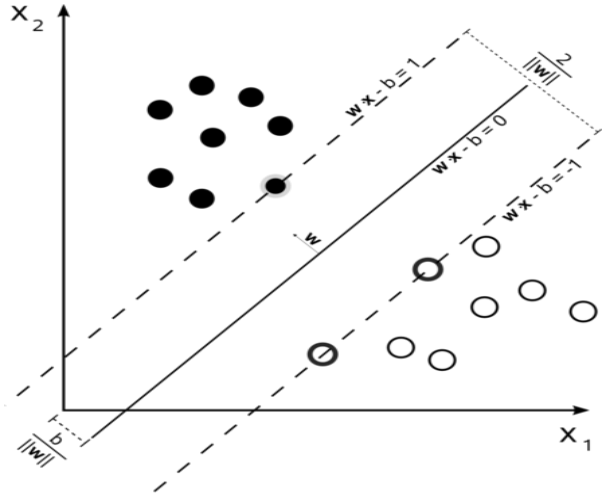


Fig 5: Example for an SVM trained with samples from two classes

IV. RESULTS AND DISCUSSION

The algorithm was implemented in MATLAB software and the experimental results are shown in this section. Figure 6 shows the sample images of Telugu words taken from scanned documents. Notice the difference in the scale and noise of these images. Some of these images also have noise introduced in them. The noise type being introduced in these images are salt & pepper noise and speckle noise. As these two noises are more difficult to remove.

Fig 6: Sample Images

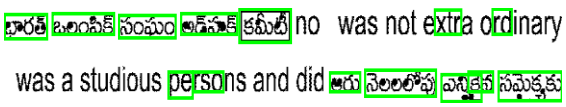
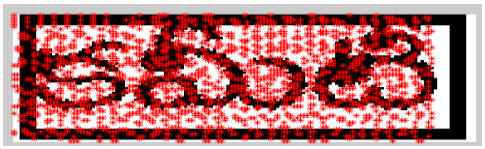
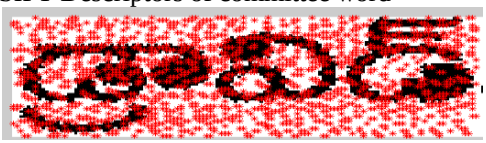


Fig 7: Segmentation of the words

The proposed method was implemented in MATLAB and the experiment was conducted by varying the size of BOW dictionary. The first step after deskewing of the document is calculating the SIFT descriptors and storing the database. The default parameters for calculating SIFT features used were as follows – patch size =16, gridSpacing=8. In Figure 8, SIFT features of the word 'committee' and 'dhravid' are shown in fig 8.



(a) SIFT Descriptors of committee word



(b) SIFT Descriptors of Dhravid word

Fig 8: SIFT Descriptors of (a) Committee and (b) Dhravid

Figure 9 shows SIFT keypoints of the word Dhravid with patchsize=16, gridspacing=16. It can be seen that increasing the gridSpacing decreases the performance of extracting SIFT features.

Fig 9: SIFT keypoint of Dhravid with gridspacing=16, patchsize=16

Figure 10 shows the result of feature extraction with the patch size being set to 32 instead of 16.

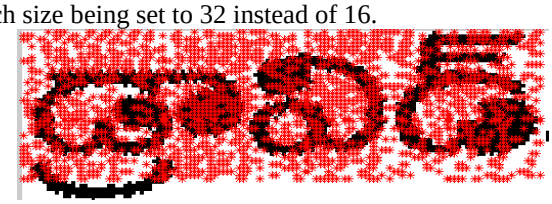


Fig 10: SIFT Descriptors of Dhravid word with patchsize=32, gridspacing=16

Based on the above figures (8, 9 and 10), it can be concluded that increasing the patchsize generates too many features while increasing the gridspacing generates too few features. The dictionary size is pre-assigned and the dictionary is built based on the basis of k-means clustering. The maximum number of iterations for clustering was taken to be 0.009. Based on the error threshold in the k-means clustering, the rate of convergence will change. The table given below shows the rate of convergence for codeword generation.

Table 1: Rate of Convergence

Error Threshold	No. of iterations
0.002	7
0.009	9
0.0009	12

Finally, 0.0009 was selected as the final error threshold. The features are then assigned to the dictionary based on their Euclidean distances. A spatial pyramid of these words is created by using the histogram of these bags of words.

The classification was done based on SVM and some of the results are presented here in figure 11.

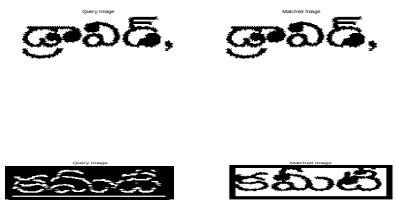


Fig 11: Matching of the word (a) Dravid and (b) Committee

Based on different kernels and vocabulary size, different accuracy results based on k-fold estimation are presented here in Figure 11. It is evident that the dictionary size has a great effect on the classification accuracy. Another intriguing effect comes from by comparing two different kernels of SVM. For reduced vocabulary sizes, the RBF kernel has a reasonable better performance in comparison to the linear one, but this merit is turned around, once the peak performance is reached.

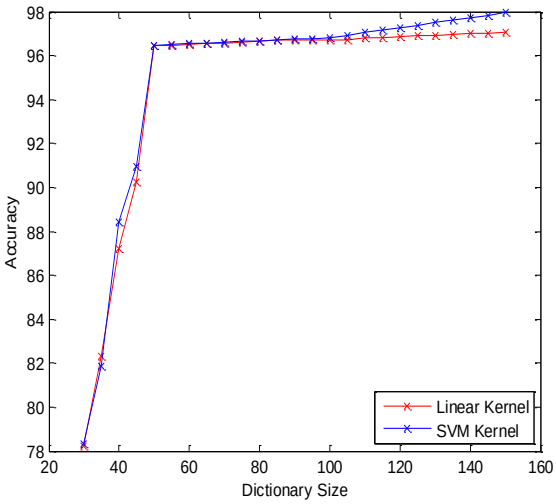


Fig 12: The classification performance at different vocabulary sizes

Table 2: Accuracy based on different kernels and vocabulary sizes

Dictionary Size	Linear Kernel	RBF Kernel
30	78.23	78.35
50	96.47	96.55
100	96.71	96.82
150	97.07	97.36

This concludes that the visual words in a reduced dictionary size are highly related to each other, yet turn out to be more independent and as the size is increased the linear separability also improves. When the visual words are independent, kernels that utilize inter-feature correlations (e.g., RBF) have no evident superiority over linear kernels and may perform inadequately because of over fitting. Based on Table II, it can be justified that irrespective of the dictionary size, the RBF kernel simply outperforms the linear kernel.

V. CONCLUSION

Bag-of-visual-word is an efficient way of representing images for the purpose of classification as compared to the other methods of image representation. In this paper, we have applied this representation to identify Telugu words taken from noisy document images. In order to compensate for rotation, we have implemented de-skewing in the preprocessing stage of the proposed algorithm. The relationship between visual words in images and words in documents opens up opportunities for redesigning techniques of image classification.

REFERENCES

1. Digital library of india. http://dli.iiit.ac.in/.
2. The universal library. http://www.uliborg.
3. D. Bainbridge, J. Thompson, and I. H. Witten, "Assembling and enriching digital library collections", In Proc. of Joint Conference on Digital Libraries, Houston, TX, USA, pp. 323–334, Jun. 2003.
4. G. Nagy, "Twenty years of document image analysis in PAMI", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 38-62, Jan. 2000.
5. W. Saffady. Introduction to Automation for Librarians. 4th edition, 1999.
6. Chanda, S., Terrades, O.R., and Pal, U.: SVM Based Scheme for Thai and English script identification. In: 9th International Conference on Document Analysis and Recognition, 1: 551-555.
7. B. B. Chaudhuri, U. Pal and M. Mitra,"Automatic recognition of Oriya script", Sadhana, vol. 27, no. 1, pp. 23-34, Feb. 2002.
8. Y. H. Tay, M. Khalid, R. Yusof and C. V. Gaudin, "Offline cursive handwriting recognition system based on hybird Markov model and neural networks", In Proc. of International Symposium on Computational Intelligence in Robotics and Automation, Aug. 2003, Kobe, Japan.
9. C. V. Jawahar, M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran, "A Bilingual OCR for Hindi-Telugu Documents and its Applications", In: Proc. of 7th International Conference on Document Analysis and Recognition, pp. 1-5, Sep. 2003.
10. N. S. Rani and T. Vasudev, "A Generic Line Elimination Methodology using Circular Masks for Printed and Handwritten Document Images", In Proc. of International Conference on Emerging Research in Computing, Information, Communication and Applications, vol. 3, no. 1, pp. 589-594, 2014.
11. R. Singh and M. Kaur, "OCR for Telugu script using backpropagation-basedclassifier", International Journal of Information Technology and Knowledge Management, vol. 2, no. 2, pp. 639-643, Jul-Dec 2010.
12. S. V. Patgar, T. Vasudev and S. Murali, "A system for detection of fabrication in photocopy document", 2nd International Conference on Computer Science and Engineering, Aug. 2015.
13. T. Q. Phan, P. Shivakumara, Z. Ding, S. Lu and C. L. Tan, 2011. Video Script Identification based on Text Lines, In Proc. ICDAR
14. P. B. Pati and A. G. Ramakrishnan. 2008. Word level multi-script identification, Pattern Recognition Letters.
15. L. Li and C. L. Tan, 2008. Script Identification of Camera-based Images, In Proc. ICPR.
16. Farhad M. M.; Nafiul Hossain S.M.; Khan A.S.; Islam A.,"An efficient Optical Character Recognition algorithm using artificial neural network by curvature properties of characters," in 2014 International Conference on Informatics, Electronics & Vision (ICIEV), vol., no., pp.1-5, 23-24 May 2014
17. Fataicha, Y.; Cheriet, M.; Nie, J.Y.; Suen, C.Y., "Information Retrieval Based on OCR Errors in Scanned Documents," in Conference on Computer Vision and Pattern Recognition Workshop(CVPRW). vol.3, no., pp.25-25, 16-22 June 2003
18. T.Rath, R. Manmatha, and V. Lavrenko.,"A search engine for historical manuscript images," In Proceedings of SIGIR04,pp. 297 -304, 2004.
19. Zhiyong Ye; Yijian Pei; Jihong Shi, "An Improved Algorithm for Harris Corner Detection," in 2nd International Congress on Image and Signal Processing(CISP '09), vol., no., pp.1-4, 17-19 Oct. 2009
20. D. Lowe, "Distinctive image features from scale-invariant keypoints," IJCV 60 (2) (2004) 91110