2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

A Study on Word-Level Multi-script Identification from Video Frames

Nabin Sharma*, Umapada Pal†, Michael Blumenstein*
*School of Information and Communication Technology, Griffith University, Australia
Email: {nabin.sharma, m.blumenstein}@griffith.edu.au
†Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India
Email: umapada@isical.ac.in

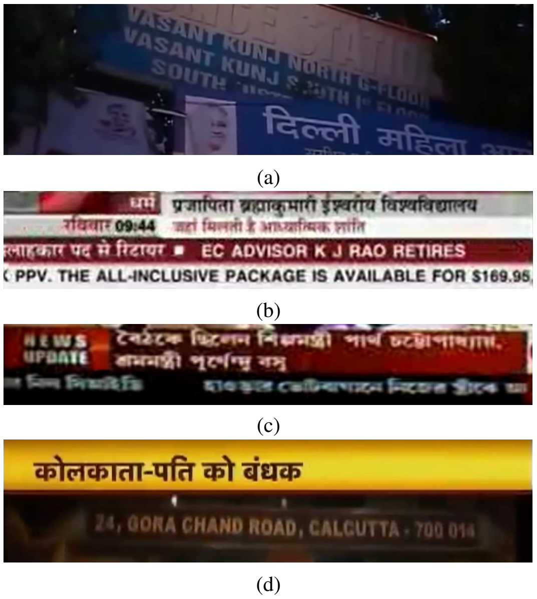
***Abstract*—The presence of multiple scripts in multi-lingual document images makes Optical Character Recognition (OCR) of such documents a challenging task. Due to the unavailability of a single OCR system which can handle multiple scripts, script identification becomes an essential step for choosing the appropriate OCR. Although, there are various techniques available for script identification from handwritten and printed documents having simple backgrounds, however script identification from video frames has been seldom explored. Video frames are coloured and suffer from low resolution, blur, complex background and noise to mention a few, which makes the script identification process a challenging task. This paper presents a study of various combinations of features and classifiers to explore whether the traditional script identification techniques can be applied to video frames. A texture based feature namely, Local Binary Pattern (LBP), Gradient based features namely, Histogram of Oriented Gradient (HoG) and Gradient Local Auto-Correlation (GLAC) were used in the study. Combination of the features with SVMs and ANNs where used for classification. Three popular scripts, namely English, Bengali and Hindi were considered in the present study. Due to the inherent problems with the video, a super resolution technique was applied as a pre-processing step. Experiments show that the GLAC feature has performed better than the other features, and an accuracy of 94.25% was achieved when testing on 1271 words from three different scripts. The study also reveals that gradient features are more suitable for script identification than the texture features when using traditional script identification techniques on video frames.**

Keywords: Video document analysis, Script identification, Word segmentation, OCR.

I. Introduction

India is a multi-lingual and multi-script country where the use of multiple scripts is quite common for information communication through news and advertisement videos transmitted across various television channels. The massive information explosion across multiple communication channels creates a very large database of videos, which makes indexing an essential task for effective management of the database. Thus, text present in the video plays an important role in automatic video indexing and retrieval. Hence, OCR of the multi-lingual video text is essential. Due to the unavailability of a universal OCR to recognize the multi-lingual text, script identification followed by the use of appropriate OCR is a legitimate approach to recognizing the text.

The research on script identification to date primarily focuses on processing scanned documents with simple backgrounds and good resolution required for OCR. Whereas the difficulties involved in script identification from video frames include low resolution, blur, complex backgrounds, multiple font types and size and orientation of the text [2], [3]. Samples of video frames having text written in multiple scripts are shown in Figure 1. Figure 1(a) is an example of a video frame having text written in English and Hindi with different orientations, fonts, and size. Figure 1(b, c) are examples of video frames having text in low resolution and blur. Figure 1(b) has text written in Hindi and English in a single text line. Figure 1(c) is an example of a video frame having text written in Bengali (Bangla) and English in a single text line. Figure 1(d) is an example of a video frame having both graphics and scene text written in Hindi and English, respectively. The English text line has little blur compared to the Hindi text which is much clearer. Figure 1 itself explains the necessity of script identification and the challenges involved when video frames are considered. An important characteristic of multilingual videos in India is that the text is generally written in two scripts, where the first script is English (Roman) and the other one is a regional language.


VASANT KUNJ NORTH G-FLOOR
VASANT KUNJ SOUTH
दिल्ली महिला
प्रजापिता ब्रह्माकुमारी ईश्वरीय विश्वविद्यालय
रविवार 09:44
EC ADVISOR K J RAO RETIRES
PPV. THE ALL-INCLUSIVE PACKAGE IS AVAILABLE FOR $169.95
NEWS UPDATE
कोलकाता-पति को बंधक
24, GORA CHAND ROAD, CALCUTTA - 700 014

(a)

(b)

(c)

(d)

Fig. 1: Samples of video frame having text in multiple scripts

Script identification from video frames has not been explored much as compared to traditional scanned documents.

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

Recently, a few papers [4], [5], [6] have been published, which focus on the video script identification problem. Sharma et. al [4] presented a study on word-wise script identification from video frames using three different features namely, Zernike moment, Gabor and 400-dimensional gradient. They used SVMs for classification. The study established that traditional script identification techniques can be applied to video frames provided appropriate pre-processing technique are applied to the video frames to overcome the problems with video. Zhao et al. [5] on the other hand proposed Spatial-Gradient-Features at the block level to identify six different scripts. The method considers text lines extracted from the video frames for the experiments, assuming that a video frame contains text written in a single script. Six different scripts were considered in the work and an average classification rate of 82.1% was reported on a dataset of 770 frames. Phan et al. [6] also proposed a line-wise script identification technique based on the smoothness and cursiveness of the lines. A video text line was horizontally divided into five equal zones to study the smoothness and cursiveness of the upper and lower lines for script identification. English, Chinese and Tamil script pairs were considered in their experiments. Li and Tan [14] proposed a statistical script identification approach from camera-based images.

There are many methods [1], [9], [8], [12], [7] available for script identification from scanned documents having simple backgrounds. A review of various script identification techniques used for script identification at the page, line and word levels, was presented by Ghosh et al. [1]. The various techniques can be classified into two broad categories, namely: structure-based and visual appearance-based methods. The review [1] shows that the methods used for traditional scanned document can be used for camera-based documents even though the former have much better resolution than the video frames, and the latter suffer from issues such as low resolution, and complex backgrounds, to mention a few. A two-stage approach based word-wise script identification technique was proposed by Chanda et al. [9]. In the first stage, a high speed identification method of scripts in noisy environment was used. The second stage processes the samples where a low recognition confidence was achieved. Finally they used a majority voting-based method to identify the script. Two different features namely, 64-dimensional chain code histogram and 400-dimensional gradient features were used in the first and second stages, respectively. English, Devanagari and Bengali scripts were considered for the experiments. The study presented by Pati and Ramakrishna et al. [8] revealed that the use of Gabor features with nearest neighbor or SVM classifiers gave a better performance for word-level multi-script identification. A combination of discrete cosine transform (DCT) features with SVMs, nearest neighbor and linear discriminant classifiers were also evaluated in their study. The authors [8] used a dataset comprising of images with simple backgrounds for their experiments.

Although there are works on line-wise video script identification, to the best of our knowledge there is only one work [4] reported in the literature on word-wise script identification from video. In this paper, a study of word-wise script identification techniques from video is presented considering Indian languages. The three most popular scripts in India namely, English, Bengali and Hindi (Devanagari) were considered for experimentation. Considering words for script identification rather than a complete text line allows the identification of the words written in different scripts, which in turn help in better OCR of the complete text line written in multiple scripts. This is an important advantage of considering words to identify scripts. Our previous study [4] revealed that the use of appropriate pre-processing techniques on the video frame is essential in order to use traditional techniques for video script identification. The present study attempts to investigate the type of features more suitable for video script identification considering the inherent problems with video, which is the main contribution. Hence, a comparison of texture-based features with gradient-based features is performed. A very popular texture-based feature namely, Local Binary Pattern (LBP) and two gradient-based features namely, Histogram of Oriented Gradient (HoG) and Gradient Local Auto-Correlation (GLAC) were used in the present study. Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs) were used for classification. As mentioned in our previous study [4] pre-processing does help in improving the video script identification accuracy, but the choice of features is also equally important. Hence, the features used in the present study were carefully selected based on their ability to provide better randomness and description of the structural differences of the scripts, which will increase the overall accuracy.

Rest of the paper is organized as follows. The pre-processing technique used is discussed in Section II. Section III presents a brief description of the feature extraction techniques used in the present study. In Section IV, the details of SVM and ANN classifiers are discussed. Experimental results and a discussion are presented in Section V. Section V also provides an analysis of the errorneous results. Section VI concludes the paper providing future directions for video script identification.

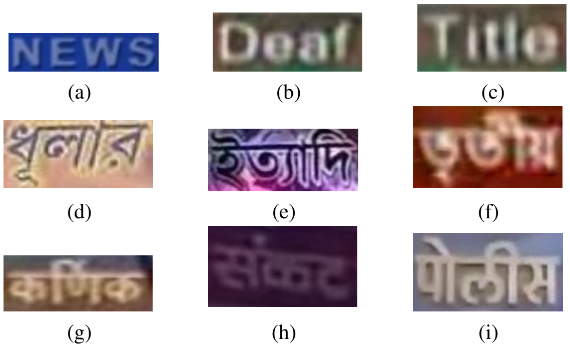


Fig. 2: Sample video word images of English (1st Row), Bengali (2nd Row) and Hindi (3rd Row) scripts.

II. Pre-processing

The text lines from the video frames were detected using [11]. The words were segmentation from the text lines using our word segmentation technique [10] and were used as input for our experiments. A few samples of segmented word images from video frames for the three scripts are shown in Figure 2. The images shown in Figure 2 reveals that the text extracted from the video frames suffers from low resolution, blur, and complex backgrounds, to mention a few issues.

Our study in [4] showed that super resolution techniques resulted in better accuracy. Hence we used the super resolution technique for pre-processing the words to get better resolution images for further processing. A single level of super resolution images were used for our experiments. The resolution of the word image was increased by 1.5% using a cubic interpolation method [13]. Cubic interpolation was chosen because it creates better images preserving the shape of the original word images.

III. Feature extraction techniques

Three feature extraction techniques were considered for the present study. One texture-based feature namely, Local Binary Pattern (LBP) and two gradient-based features namely, Histogram of Oriented Gradients (HoG) and Gradient Local Auto-Correlation (GLAC) were used. A brief description of the feature extraction techniques are discussed below.

A. Local Binary Pattern (LBP)

Local Binary Pattern (LBP) [15] is an efficient texture operator which labels each pixel of an image by thresholding their neighbours. The idea behind the LBP operator is to describe the image textures using two measures namely, local spatial patterns and the gray scale contrast of its strenght.

We considered the original version of the LBP operator [15] which forms labels of image pixels by thresholding the 3×3 neighbourhood of each pixel with the centre pixel value and the result is considered as a binary number. As the neighbourhood of the centre pixel has 8 pixels, $2^8 = 256$ different labels can be obtained and used as a texture descriptor. A histogram is then computed over the cells, and forms the feature vector.

The basic $LBP_{P,R}$ operator is defined as follows,

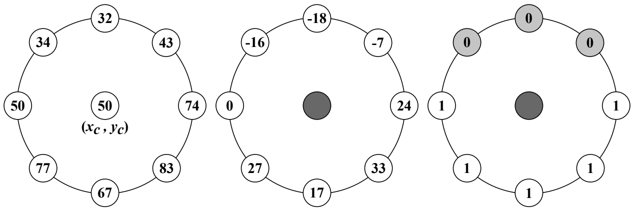
$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} S(g_p - g_c)2^P \qquad (1)$$

where,

$$S(x) = \begin{cases} 1, if x >= 0 \\ 0, otherwise \end{cases}$$

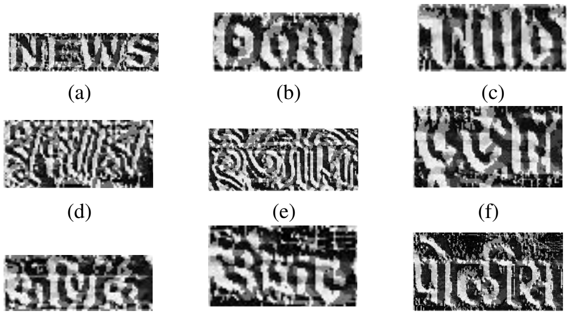
S(x) is a thresholding function, (x_c, y_c) is the centre pixel in the 8 pixel neighbourhood, g_c is the gray level of the centre pixel and g_p denotes the gray value of a sampling point in an equally spaced circular neighbourhood of P sampling points and radius R around the point (x_c, y_c). An illustration of LBP computation is shown in Figure 3. Figure 4 shows the LBP images correspoding to the sample video word images shown in Figure 2.

LBP was chosen for the present study because of its ability to describe the local spatial pattern, which required discriminate between the structurally similar scripts.



(a) Sample pixel neighbourhood (b) Difference result (c) Thresholding result

Fig. 3: An example of LBP computation



(a) (b) (c)
(d) (e) (f)
(g) (h) (i)

Fig. 4: LBP Images of the corresponding video word images shown in Figure 2 for the three scripts.

B. Histogram of Oriented Gradients (HoG)

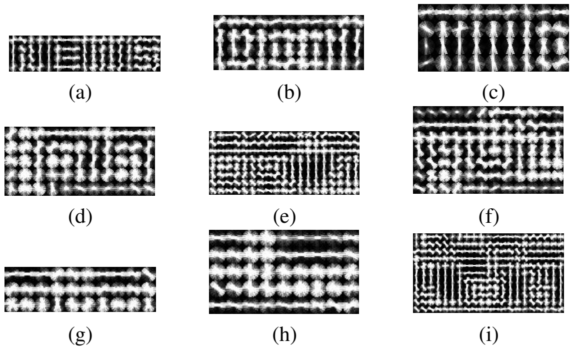
Histogram of Oriented Gradients (HoG) [16] is a robust feature descriptor commonly used in computer vision and image processing for object detection. Dalal and Triggs [16] first described the HoG descriptors and primarily focused on pedestian detection in static images. The basic idea behind the HoG descriptor is that the shape and appearence of the object within an image can be described by the intensity gradient distribution or the edge directions.

The HoG descriptors are typically computed by dividing an image into small spatial regions called 'cells'. A histogram of the gradient direction of the pixels within the cells is computed. The histogram bins/channels are evently spaced over 0° to 180° or 0° to 360° based on the usage of signed or unsigned gradient values. Combining the histogram of all the cells produces the descriptors. For improving the accuracy the local histograms can be contrast-normalized [16]. More information about the HoG descriptor can be found in [16].

For our study the HoG feature suits the problem well because it operates on the localized cells and it is capable of describing the shape and appearance of the object, which is the word in the present context. Figure 5 shows the HoG images correspoding to the sample video word images shown in Figure 2.

C. Gradient Local Auto-Correlation (GLAC)

Gradient Local Auto-Correlation (GLAC) was proposed by Kobayashi and Otsu [17]. It utilizes the spatial and the orientational auto-correlations of local gradients for feature extraction. The features not only capture the information about the gradients but also the curvature of the image surface, and are described in terms of both magnitude and orientation.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 5: HoG Images of the corresponding video word images shown in Figure 2 for the three scripts.

GLAC can be viewed as an extension of 1st order statistics (i.e. histograms) to the 2nd order statistics (that is the auto-correlations).

Detailed information about GLAC features can be found in [17]. The ability of GLAC features to describe the curvature of the image surface, in addition to the gradient information, inspired us to consider it in our study.

IV. Classifiers

We considered Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs) for classification of the three scripts. A brief description of the classifier are given below.

A. Support Vector Machine (SVM)

Given a training database of M data: $\{x_m | m = 1, ..., M\}$, the linear SVM classifier is then defined as:

$$f(x) = \sum_j \alpha_j y_j x_j \cdot x + b \qquad (2)$$

where, x_j are the set of support vectors, y_j is the set of class labels {+1, -1} and the parameters a_j and b have been determined by solving a quadratic problem [18]. The linear SVM can be extended to various non-linear variants; details can be found in [18], [19]. In our experiments Gaussian kernel SVM outperformed linear and other non-linear SVM kernels. The Gaussian kernel is of the form:

$$k(x, y) = exp\frac{-||x-y||^2}{2\sigma^2} \qquad (3)$$

We noticed from the initial experiments that the Gaussian kernel gave the highest accuracy when the value of its gamma parameter $(1/2\sigma^2)$ was varied between 1.0 and 5.0 for the three different features and the penalty multiplier parameter was set to 1. LibSVM [20] was used to conduct the SVM classification experiments.

B. Artificial Neural Networks (ANNs)

In this study, feed-forward Multi-Layered Perceptrons (MLPs) trained with the resilient backpropagation (BP) algorithm were used. For experimental purposes, the architectures were modified varying the number of inputs and the hidden units. The number of output units were fixed to three as three scripts were considered in the present study. The number of input units varied because three different feature having different feature dimension were considered in the present study.

The number of hidden units investigated during ANN training was experimentally set from 8 upto 30 hidden units. The number of iterations set for training was increased from 1000 upto 3000. All the ANNs were trained with a learning rate of 0.1 and a momentum rate of 0.1.

V. Experimental Results

This section presents the experimental results obtained using the various combinations of the three features, as well as the SVM and ANN classifiers. In order to study the performance of the features and classifiers, a video word dataset was created, as there is no standard dataset available. Test portion of a video frames were extracted using our video text detection algorithm [11] and the words were later segmented using our word segmentation technique [10]. A dataset of 1271 words was created after extraction of word from the video text lines. The dataset comprised of 430 Hindi, 410 English and 431 Bengali words. The results obtained using the combinations of features and classifiers are reported in Tables I, II, III, IV, V, VI and VII. For all the experiments we used a five fold cross validation technique to compute the script identification accuracy. The reason for using cross validation is that it provided unbaised results over the complete dataset.

The various experiments conducted in the study included the perfomance evaluation of :

- LBP features with SVM and ANN classifiers,
- HoG features with SVMand ANN classifiers,
- GLAC features with SVM and ANN classifiers.

Additionally, we also conducted experiments on Long-words (words having four or more characters) and Short words (word having three or less characters). We evaluated the performance of HoG and GLAC features with SVM and ANN classifiers on both Long and Short words. This experiments revealed the discriminative capacity and robustness of the features when applied on the same dataset and their impact on the accuracy.

A. Experimental settings

The parameter settings considered for each of the feature extraction techniques used in the present study are given below.

1) LBP feature: as mentioned earlier, the basic LBP was considered in our study, with 8 neighbours the feature dimensions of the LBP feature vector was 256.
2) HoG feature: The block size considered was 5. That is an image was divided in 5×5 blocks. The gradient orientation was quantized into 16 directions/bins. Thus, the feature dimensions of a HoG feature vector was $5 \times 5 \times 16 = 400$.
3) GLAC feature: the Roberts filter was used for gradient computation and the number of orientation bins was set to 9. The other baseline parameter settings as given in [17] were considered.

B. *Performance using LBP features*

Experiments using LBP with SVMs and ANNs resulted in a comparatively lower accuracy when compared to the accuracy obtained using HoG and GLAC. The accuracy obtained using LBP with both SVMs and ANNs is given in Table I. The confusion matrix presented in Table I for SVMs shows that 92.84% accuracy was obtained, whereas using ANNs the accuracy was 85.29%. The confusion matrix in Table I also reveals that highest confusion of about 9.51% and 11.37% was between Bengali and Hindi using SVM and ANN classifiers, respectively. The overall accuracy obtained using LBP and SVMs was much better than ANNs. The two texture-based features namely, Zernike moments and the Gabor filter used in our previous study [4] also did not perform well compared to the gradient feature. Although, video frames suffers from low resolution, blur, complex backgrounds etc, LBP still performed much better with SVM than Zernike moments and the Gabor filter.

Another reason for the overall lower performance of texture features is, they tend to describes the texture of scripts rather than the overall structure/shape. Bengali and Hindi scripts are confused mostly because of their structural similarity.

TABLE I: Confusion matrix for script identification using LBP feature with SVM and ANN classifiers

Classifier	SVM			ANN		
Scripts	English	Bengali	Hindi	English	Bengali	Hindi
English	415	11	5	368	33	27
Bengali	20	370	41	41	341	49
Hindi	12	23	395	16	39	375
Accuracy	92.84%			85.29%		

C. *Performance using HoG features*

The results obtained using the combination of HoG features with SVMs and ANNs are reported in Table II. Table II shows that HoG with SVMs performed better than HoG with ANNs and LBP. A 93.78% accuracy was obtained using HoG and SVMs, whereas the combination of HoG and ANNs resulted in 90% accuracy. The table show that highest confusion occured for Bengali and Hindi, where 12.29% of the Bengali words were confused with Hindi script using ANN.

The results obtained with HoG are much more consistent than that of LBP, where the difference between the accuracy of LBP with SVMs is considerably much higher than LBP and ANNs. This gives an indication that the texture features were not able to describe the shape/structure of the scripts properly. Conversely, HoG features performed much better than LBP. One possible reason could be the usage of edge directions to describe the shape/structure of the scripts.

TABLE II: Confusion matrix for script identification using HoG feature with SVM and ANN classifiers

Classifier	SVM			ANN		
Scripts	English	Bengali	Hindi	English	Bengali	Hindi
English	412	14	5	399	27	5
Bengali	24	377	30	23	355	53
Hindi	6	21	403	14	27	390
Accuracy	93.78%			90.00%		

D. *Performance using GLAC features*

Table III presents the confusion matrix obtained using the combination of GLAC with SVMs and ANNs. The highest accuracy of 94.25% was achieved using GLAC and SVMs. Also the performace with ANNs was also comparable with an accuracy of 93.55%. The results obtained using GLAC features are much better and more consistent than HoG and LBP, with very less difference between the performance of SVMs and ANNs. This also reveals that GLAC features are much more robust than HoG and LBP. The reason behind the better performance using GLAC features is that it is uses gradients and curvature of the image surface for feature description. The use of curvature information provides better information about the shape/structure of the scripts which in turn helped the classifiers to model the script patterns more accurately.

Also, with GLAC features, Bengali script was confused mostly with Hindi having an error of about 8.58% when using SVMs. The possible reason for confusion are discussed in the error analysis subsection of section V.

TABLE III: Confusion matrix for script identification using GLAC feature with SVM and ANN classifiers

Classifier	SVM			ANN		
Scripts	English	Bengali	Hindi	English	Bengali	Hindi
English	416	9	6	412	12	7
Bengali	17	377	37	24	375	32
Hindi	6	20	405	6	22	402
Accuracy	94.25%			93.55%		

E. *Experiments on Long and Short words*

Another set of experiments were conducted to understand the impact of the number of characters in the word, towards the overall script identification accuracy. For this experiment the dataset was divided into two subsets, consisting of long words (words having four or more characters) and short words (words have three or less characters). The long word dataset formed thus comprised of 325 English, 236 Bengali and 197 Hindi words. Whereas, the short word dataset consisted of 107, 174 and 233 words for English, Bengali and Hindi scripts, respectively. For this experiment HoG and GLAC features were considered as they performed better than LBP.

The results obtained on the long word dataset is presented in the Tables IV and V for HoG and GLAC features, respectively. Table V shows that the highest accuracy of 96.04% was acheived using GLAC with SVMs, whereas the GLAC features with ANNs produced a 95.24% accuracy. It can be clearly seen that the accuracy obtained for the long words using GLAC features is 1.79% (i.e. 96.04 - 94.25) more than the accuracy obtained for the combined dataset. Thus, confirming the observation from our previous study [4], that due to the presence of more characters in the long words, more script specific information is available, which resulted in the increase of the accuracy.

The script identification accuracy obtained on short words is reported in the Tables VI and VII using HoG and GLAC features, respectively. The highest accuracy of 94.16% was obtained using GLAC and SVMs on the short word dataset. The result is slightly (0.09%) less than the accuracy obtained on the combined dataset using GLAC and SVMs. The difference is more significant when we examine the results obtain using HoG on short words and the combined datasets. The result obtained using HoG and SVMs on the short word dataset is 1.57% lower than the accuracy obtained on the combined dataset. The accuracy (87.54%) obtained using HoG and ANNs is still much lower (2.46%) than that of the combined dataset. Thus, when there are fewer characters in the word, it results in more confusion and misclassifications.

TABLE IV: Confusion matrix for script identification using HoG feature with SVM and ANN classifiers on long words

Classifier	SVM			ANN		
Scripts	English	Bengali	Hindi	English	Bengali	Hindi
English	320	3	1	305	14	5
Bengali	16	202	18	6	212	18
Hindi	4	13	180	5	18	174
Accuracy	92.75%			91.23%		

TABLE V: Confusion matrix for script identification using GLAC feature with SVM and ANN classifiers on long words

Classifier	SVM			ANN		
Scripts	English	Bengali	Hindi	English	Bengali	Hindi
English	315	5	4	314	6	4
Bengali	6	225	5	8	218	10
Hindi	4	6	187	3	5	189
Accuracy	96.04%			95.24%		

TABLE VI: Confusion matrix for script identification using HoG feature with SVM and ANN classifiers on short words

Classifier	SVM			ANN		
Scripts	English	Bengali	Hindi	English	Bengali	Hindi
English	100	3	4	92	12	3
Bengali	4	152	18	5	146	23
Hindi	3	8	222	6	15	212
Accuracy	92.21%			87.54%		

TABLE VII: Confusion matrix for script identification using GLAC feature with SVM and ANN classifiers on short words

Classifier	SVM			ANN		
Scripts	English	Bengali	Hindi	English	Bengali	Hindi
English	101	3	3	100	4	3
Bengali	1	161	12	5	154	15
Hindi	1	10	222	3	12	218
Accuracy	94.16%			91.82%		

F. Error Analysis

The experimental results show that Bengali script was mostly confused with Hindi. A further investigation was done to understand the reasons behind the mis-classification. The low resolution and blurred images of the words of the three scripts were seperated to form a dataset. Another dataset was formed by selecting the sharp and better resolution images of the three scripts. In total the low resolution and blurred image dataset was formed using 235 word images comprising of 71 English, 92 Bengali and 73 Hindi words. Whereas, the better resolution and sharp image dataset comprised of 1035 words, having 360 English, 318 Bengali and 357 Hindi words.

GLAC features were extracted for the images in both the datasets and were tested using SVM. Only SVM was considered for this experiement because it performed well in the earlier experiments. The accuracy obtained from the experiments are presented in table VIII and IX. A 5-fold cross validation technique was used to compute the accuracy.

TABLE VIII: Confusion matrix for script identification using GLAC feature with SVM for error analysis using good resolution word images

Scripts	English	Bengali	Hindi
English	351	5	4
Bengali	5	295	18
Hindi	2	7	348
Accuracy	96.14%		

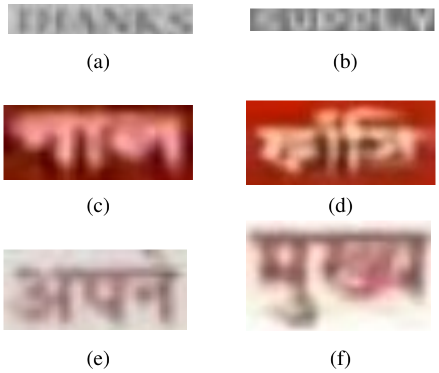
TABLE IX: Confusion matrix for script identification using GLAC feature with SVM for error analysis using blur and low resolution word images

Scripts	English	Bengali	Hindi
English	67	1	3
Bengali	4	80	8
Hindi	1	4	68
Accuracy	91.49%		

An accuracy of 96.14% was obtained when only good resolution and sharp images of the words were considered. Whereas, the accuracy decreased to 91.49% when the low resolution and blurred image where considered. This experiments confirms that apart from the structural similarities between the scripts, low resolution and blur were the major reasons behind the confusion between the scripts. Due to low resolution and blur, the Bengali script looked like Hindi and hence resulted in the maximum misclassification. A few sample of errorneous word images are shown in Figure 6.

The Bengali word images in Figure 6 (c, d) were misclassified as Hindi because of the very low resolution, blur and the fewer number of characters. The English word images in Figure 6 (a, b) were classified as Hindi and Bengali, respectively. The English word images also having very low resolution which confuses even human beings at the first instance, hence resulting in the misclassification. The Hindi word images shown in Figure 6 (e, f) were misclassified as Bengali: low-resolution and blur were the main reasons for the same.

The percentage of errors which occurred in different experiments are presented in Table X. Table X shows that Bengali script has the highest error rate of 13.03% and 7.23% in the experiments with low resolution and blurred images, and high resolution and sharp images, respectively. For Hindi script, 6.85% error occurred when low resolution and blurred images



(a) (b)

(c) (d)

(e) (f)

Fig. 6: Some errorneous samples of three scripts which were misclassified.(a, b) English word images; (c, d) Bengali word images; (e, f) Hindi word images

TABLE X: Error distribution in different experiments

Scripts	Error (in %)		
	Blur and low resolution images	Sharp and better resolution images	Complete Dataset
English	5.71	2.5	3.65
Bengali	13.03	7.23	12.53
Hindi	6.85	2.52	6.05

were considered, whereas, the error reduced to 2.52% for high resolution images. English script confused lesser than Bengali and Hindi. Considering the low resolution and blurred images in English script, 5.71% error rate was observed, whereas, 2.5% error occurred with high resolution images. The error obtained on complete dataset, which is a mixture of both low resolution and high resolution images, was also computed to understand how much error does the low resoluton and blurred images contributed to the overall accuracy. The error obtained on the complete dataset also had the same characteristics. For the whole dataset, 3.65%, 12.53% and 6.05% errors were observed for English, Bengali and Hindi scripts, respectively. Although, pre-processing the words using super-resolution techniques indeed increased the overall accuracy, but the very low resolution and blur images still confused with other scripts.

VI. Conclusions

This paper presented a study of various techniques for word-wise video script identification. A comparative study of the combination of texture and gradient-based features with two classifiers was presented in the paper. SVMs and ANNs were used for the classification experiments. A large dataset with complex backgrounds was used for the experiments and the results obtained were promising. The experiments show that the combination of GLAC features with SVMs performed better than the others and 94.25% accuracy was obtained. The study reveals that gradient-based feature are better than texture-based. It may be noted that GLAC features particularly performed better with both SVMs and ANNs. In general, features which can describe the shape/structure of the script more robustly can be considered for script identification in video. The texture feature did not perform well because it only described the texture rather than the structure of the scripts. Future research plans include to study classifier fusion and feature-fusion based techniques on more Indian scripts in order to create a more robust system capable of handling multiple scripts, accurately.

References

[1] D. Ghosh, T. Dube and A. P. Shivaprasad, *Script Recognition- Review*, IEEE Transactions on PAMI, Vol-34, pp. 2142-2161, 2010.

[2] N. Sharma, U. Pal, and M. Blumenstein. *Recent Advances in Video Based Document Processing: A Review*, In Proc. DAS, pp. 63-68, 2012.

[3] K. Jung, K.I. Kim and A.K. Jain, *Text information extraction in images and video: a survey*, Pattern Recognition, Vol-37, no. 5, pp. 977-997, 2004.

[4] N. Sharma, S. Chanda, U. Pal and M. Blumenstein, *Word-wise Script Identification from Video Frames*, In Proc. ICDAR, pp. 38-42, 2013.

[5] D. Zhao, P. Shivakumara, S. Lu and C. L. Tan, *New Spatial-Gradient-Features for Video Script Identification*, In Proc. DAS, pp. 38-42, 2012.

[6] T. Q. Phan, P. Shivakumara, Z. Ding, S. Lu and C. L. Tan, *Video Script Identification based on Text Lines*, In Proc. ICDAR, pp. 1240-1244, 2011.

[7] H. Ma and D. Doermann, *Word Level Script Identification for Scanned Document Images*, in Proc. SPIE Document Recognition and Retrieval XI, pp. 124-135, 2003.

[8] P. B. Pati and A. G. Ramakrishnan, *Word level multi-script identification*, Pattern Recognition Letters, pp. 1218-1229, 2008.

[9] S. Chanda, S. Pal, K. Franke and U. Pal, *Two-stage Apporach for Word-wise Script Identification*, In Proc. ICDAR, pp. 926-930, 2009.

[10] N. Sharma, P. Shivakumara, U. Pal, M. Blumenstein and C. L. Tan, *A New Method for Word Segmentation from Arbitrarily-Oriented Video Text Lines*, In Proc. DICTA, pp. 1-8, 2012.

[11] N. Sharma, P. Shivakumara, U. Pal, M. Blumenstein, C. L. Tan, *A New Method for Arbitrarily-Oriented Text Detection in Video*, In Proc. DAS, pp. 74-78, 2012.

[12] S. Jaeger, H. Ma, and D. Doermann, *Identifying Script on Word-Level with Informational Confidence*, In Proc.8th ICDAR, pp. 416-420, 2005.

[13] R. Keys, *Cubic convolution interpolation for digital image processing*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol.29, no.6, pp. 1153-1160, 1981.

[14] L. Li and C. L. Tan, *Script Identification of Camera-based Images*, In Proc. ICPR, pp. 1-4, 2008.

[15] Ojala, T., Pietikinen, M. and Menp, T. , Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, No. 7, pp. 971-987, 2002.

[16] N. Dalal and B. Triggs, *Histogram of Oriented Gradients for Human Detection*, In Proc. CVPR, vol. 1, pp. 886-893, 2005.

[17] T. Kobayashi and N. Otsu, Image Feature Extraction Using Gradient Local *Auto-correlations, Proc. European Conference on Computer Vision (ECCV)*, pp. 346-358, 2008.

[18] C. Burges, *A Tutorial on Support Vector Machines for Pattern Recognition*, Data mining and knowledge discover, 2, pp. 1-43, 1998.

[19] V. Vapnik, *The nature of statistical learning theory*, Springer-Verlag, 1995.

[20] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM: A library for support vector machines*, ACM Transactions on Intelligent Systems and Technology, vol. 2, issue. 3, pp. 27:1–27:27, 2011.