Word-wise Script Identification from Video Frames

Nabin Sharma*, Sukalpa Chanda†, Umapada Pal‡ and Michael Blumenstein*
*Griffith University, Queensland, Australia 4222. Email: {nabin.sharma, m.blumenstein}@griffith.edu.au
†Gjøvik University College, Gjøvik-2802, Norway. Email: sukalpa@ieee.org
‡CVPR Unit, Indian Statistical Institute, Kolkata, India 700108. Email: umapada@isical.ac.in

***Abstract*—Script identification is an essential step for the efficient use of the appropriate OCR in multilingual document images. There are various techniques available for script identification from printed and handwritten document images, but script identification from video frames has not been explored much. This paper presents a study of some pre-processing techniques and features for word-wise script identification from video frames. Traditional features, namely Zernike moments, Gabor and gradient, have performed well for handwritten and printed documents having simple backgrounds and adequate resolution for OCR. Video frames are mostly coloured and suffer from low resolution, blur, background noise, to mention a few. In this paper, an attempt has been made to explore whether the traditional script identification techniques can be useful in video frames. Three feature extraction techniques, namely Zernike moments, Gabor and gradient features, and SVM classifiers were considered for analyzing three popular scripts, namely English, Bengali and Hindi. Some pre-processing techniques such as super resolution and skeletonization of the original word images were used in order to overcome the inherent problems with video. Experiments show that the super resolution technique with gradient features has performed well, and an accuracy of 87.5% was achieved when testing on 896 words from three different scripts. The study also reveals that the use of proper pre-processing approaches can be helpful in applying traditional script identification techniques to video frames.**

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

I. Introduction

In a multilingual and multi-script country like India, news and advertisement videos transmitted across various television channels consist of texts written in multiple scripts. For automatic indexing of such videos it is necessary to recognize the text. Since developing a universal OCR to recognize text in multiple scripts is difficult and not available, hence the appropriate approach would be to identify the script of the text first, followed by an OCR for the recognized script. Unfortunately research on script identification to date mainly emphasizes script identification in scanned documents. Moreover, script identification in video frames imposes additional challenges such as low resolution, blur, complex backgrounds, multiple font types and size and orientation of the text [2], [3]. Samples of video frame having text written in multiple scripts are shown in the Figure 1. Figure 1(a, b) are examples of video frames having English and Hindi text present in a single text line. Figure 1(c, d) are examples of video frames having English and Bengali text. An important characteristic of multilingual videos in India is that the text is generally written in two scripts, where the first script is English and the other one is a regional language or national language Hindi.

Script identification from video frames has been explored far less as compared to traditional scanned documents. Recently, a few papers [4], [5] have been published, which address the video script identification problem. Zhao et al. [4] proposed Spatial-Gradient-Features at the block level to identify six different scripts. The method was based on the text lines extracted from video frames and assumed that a single script is present in a video frame. The authors used 770 frames of six different scripts and obtained an average classification rate of 82.1%. Line-wise script identification was proposed by Phan et al. [5]. The average angle of the upper and lower lines were used to study the smoothness and cursiveness of the lines. A text line was horizontally divided into five equal zones to study the cursiveness and smoothness of upper and lower lines of each zone to discriminate the scripts. The authors [5] considered English, Chinese and Tamil script pairs for there experiments. A statistical script identification approach from camera-based images was proposed by Li et al. [15]. There are many methods [1], [7], [6], [11], [14] available

(a) (b) (c) (d)

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

for script identification from traditional scanned documents which have simple backgrounds. Ghosh et al. [1] presented a review of the various script identification techniques at the page, line and word levels. The authors grouped the various techniques into two broad categories, namely: structure-based and visual appearance-based methods. The review indicates that the methods used for traditional scanned document can be used for camera-based documents as the former have much better resolution than the video frames, as the latter suffers from issues such as low resolution, complex backgrounds, to mention a few. Chanda et al. [7] proposed a word-wise script identification technique and used a two-stage approach. 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. A majority voting technique was used in the final stage to recognize the script. The authors considered

English, Devanagari and Bengali scripts for their experiments and used a 64-dimensional chain code histogram and 400-dimensional gradient features in the first and second stages, respectively. Pati and Ramakrishna et al. [6] showed that the combination of Gabor features with nearest neighbor or SVM classifiers gives a better performance for word-level multi-script identification. They also evaluated the combination of discrete cosine transform (DCT) features with SVM, nearest neighbor and linear discriminant classifiers in their study. The data used by [6] were images with simple backgrounds.

Although there are works on line-wise script identification, to the best of our knowledge there is no work 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 the Indian languages. The three most popular scripts in India namely, English, Bengali and Hindi (Devanagari) were considered for experimentation. As words are considered for script identification, multiple scripts present in a text line can be identified using the proposed method. A comparison of texture analysis-based features (Gabor feature, Zernike moment) is performed with gradient directional features, and SVMs are used as the classifier. In order to overcome the problems with video frames as mentioned earlier, some pre-processing techniques were also explored to study their impact on the overall accuracy.

The preprocessing techniques are discussed in the Section II. The Section III presents a brief description of the feature extraction techniques used in the present study. In Section IV, the details of SVM classifier are discussed. Experimental results and a discussion are presented in Section V. Section VI concludes the paper providing future directions toward video script identification.

II. Pre-processing techniques

The input for the proposed word-wise script identification approach are the words separated using our word segmentation technique [8] from video frames. 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, complex backgrounds, to mention a few issues. In order to minimize the effect of the inherent problems with video frames, pre-processing techniques were applied on the word images. The idea behind the pre-processing technique selection in the present study was to extract the structural features of the scripts in a better way.

Two different pre-processing techniques were used and their impact on the accuracy of script identification was analysed. The first technique used was the skeleton of the word image. Skeletons approximately resembles the structure of the text and also preserves the structural properties of scripts. The second technique used was super resolution which enhances the resolution of the images without altering the structural properties of the text. The resultant images obtained after applying the pre-processing techniques subsequently were used for feature extraction. The pre-processing techniques are described below.

Skeleton image: In-order to form the skeleton of the characters present in the words, the word image is first binarized. To binarize the gray scale word images, k-means clustering of the gray values was performed with k=2. Clustering produces two clusters, one has high gray values (C_1) and the other has low gray values (C_2). The foreground cluster was identified by considering the border rows and columns of the word image, as pixels near the boundary of the word image generally belong to the background. If the number of pixels in the border rows and columns belonging to C_1 is greater than that for the pixels belonging to C_2, then C_1 is considered as the background cluster else C_2 is considered as the background cluster. Once the background cluster is identified, a binary image is created by labeling the background and foreground cluster pixels as 0 and 1, respectively. Results after binarization of the images of Figure 2 are shown in the Figure 3. It can be seen that the foreground pixels represent the characters, and the skeletons of these characters are formed. An inbuilt Matlab function was used and the default values of the required parameters were considered. Skeleton images of the word samples given in the Figure 2 are shown in Figure 4.

Fig. 2: Sample word images of English, Bengali and Hindi scripts. First, second and third row shows the samples of English, Bengali and Hindi words.

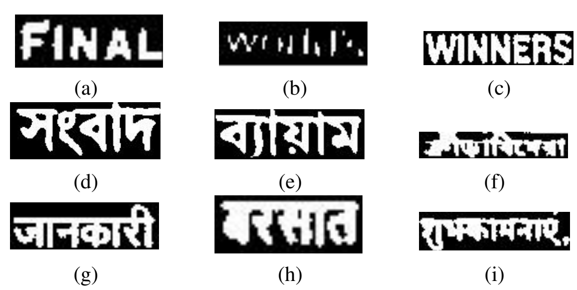

Fig. 3: Binarized images of words written in English, Bengali and Hindi scripts

Super resolution: Super resolution refers to a technique used to enhance the resolution of an image. We use a single level of super resolution images for our experiments. The resolution of the word image was increased by 1.5% using a cubic interpolation method [13]. As cubic interpolation technique produces better images and also preserves the shape of the original word images, it was chosen to created the super resolution images.

III. Feature extraction techniques

Two texture-based feature extraction techniques namely Gabor filters, Zernike moments and gradient directional

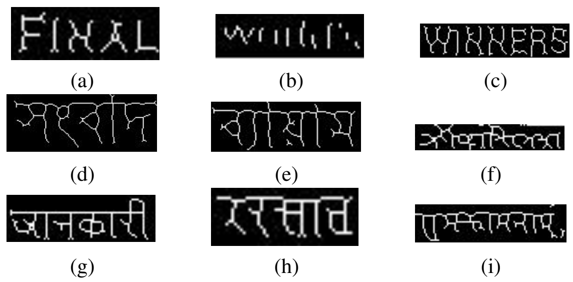

Fig. 4: Skeleton images of words written English, Bengali and Hindi scripts

feature were used in the present study. The features extraction techniques were applied to the whole word. A brief description of the feature extraction techniques are discussed below.

Gabor Filter: The spectral patterns of a document background could be quite different and therefore well-suited for texture analysis. Gabor filter-based features [12] are therefore used for this purpose. Gabor filters are capable of representing signals in both the frequency and the time domains. A two-dimensional Gabor filter in the spatial and frequency domains can be defined by the following formula:

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = exp(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}) \cos(2\pi \frac{x'}{\lambda} + \psi)$$

where

$$\begin{cases} x' = x\cos\theta + y\sin\theta, \\ y' = -x\sin\theta + y\cos\theta, \end{cases}$$

In this equation, λ represents the wavelength of the cosine factor, θ represents the orientation of the normal to the parallel stripes of a Gabor function, ψ is the phase offset, σ is the sigma of the Gaussian envelope and γ is the spatial aspect ratio, and specifies the ellipticity of the support of the Gabor function. We investigated a combination of different values of these parameters during our experiments. We achieved the best optimized results for script identification when the orientation of the Gabor filter was set to π/4, with spatial frequency set to $\sqrt{2}$, and sigma set to 2*π.
The results of size normalization and Gabor response corresponding to the input image samples taken from each of the scripts are shown in the Figure 5.

Zernike moments: Zernike moments are a class of orthogonal moments and are well suited for image representation. Zernike moments are rotation invariant in nature and can be easily constructed to an arbitrary order. Although higher order moments carry more information about an image, they are not reliable features in presence of noise. By means of empirical investigation, the optimal orders of Zernike moments were determined in the context of our problem as in [10]. The Zernike polynomials are a set of complex, orthogonal polynomials defined over the interior of a unit circle $x^2 + y^2 = 1$. Two dimensional Zernike moments can be computed using the formula:

$$A_{mn} = \frac{m+1}{\pi} \int_x \int_y f(x, y)\ V^*_{mn}(x, y)\ dx\ dy \qquad (1)$$

where $x^2 + y^2 \leq 1$ and $m - |n| = even$ and $|n| \leq m$. Here m = 0, 1, 2.... defines the order and $f(x, y)$ is the function being described and * denotes the complex conjugate. n is an integer implying the angular dependence.
For a discrete image pixel P(x,y), the integrals are changed to a summation, and the above equation gets transformed to the following:

$$A_{mn} = \frac{m+1}{\pi} \sum_x \sum_y f(x, y)\ V^*_{mn}(x, y) \qquad (2)$$

where $x^2 + y^2 \leq 1$.
For our case the idea is to map the image of the size-normalised word images to the unit disc using polar coordinates, where the centre of the image is the origin of the unit disc. Those pixels falling outside the unit disc are not used in our computation. Details about the feature can be found in [10].

Gradient directional feature: The 400-dimensional gradient directional feature [16] was used. To obtain 400-dimensional features we apply the following steps.

1) If the input image is a binary image then it is converted into a gray-scale image applying a 2x2 mean filtering 5 times.
2) The gray-scale image is normalized so that the mean gray scale becomes zero with maximum value 1.
3) Normalized image is then segmented into 9x9 blocks.
4) A Roberts filter is then applied on the image to obtain gradient image. The arc tangent of the gradient (direction of gradient) is quantized into 16 directions and the strength of the gradient is accumulated with each of the quantized direction. By strength of Gradient $f(x, y)$ we mean,
   $f(x, y) = \sqrt{(\Delta u)^2 + (\Delta v)^2}$ and by direction of gradient θ(x, y) we mean,
   θ (x, y)= $\tan^{-1} \frac{\Delta v}{\Delta u}$,
   where,
   $\Delta u = g(x + 1, y + 1) - g(x, y)$,
   Δv = $g(x + 1, y) - g(x, y + 1)$ and
   $g(x, y)$ is a gray scale at (x, y) point.
5) Histograms of the values of 16 quantized directions (with an interval of 22.5°) are computed in each of 9x9 blocks.
6) 9x9 blocks are down sampled into 5x5 by a Gaussian filter. Thus, we get 5x5x16 = 400 dimensional feature.

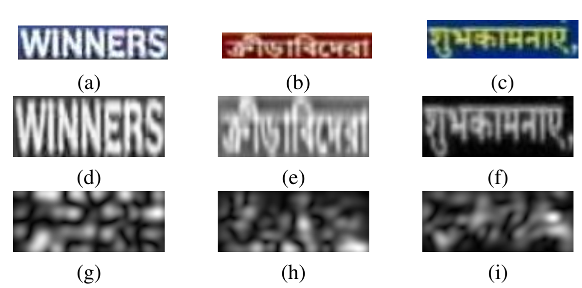

Fig. 5: Samples size normalized (d, e, f) and Gabor response images (g, h, i) corresponding to the input word images (a, b, c)

IV. Classifier

In our experiments, we have used a Support Vector Machine (SVM) as classifier. Given a training database of M data: $\{x_m | m = 1, ..., M\}$, the linear SVM classifier is then defined as:

$$f(x) = \sum_j \alpha_j y_j x_j \cdot x + b \quad (3)$$

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 has 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:

$$k(x, y) = exp\frac{-||x - y||^2}{2\sigma^2} \quad (4)$$

We noticed that Gaussian kernel gave highest accuracy when the value of its gamma parameter $(1/2\sigma^2)$ is 48.00 and the penalty multiplier parameter is set to 1.

V. Experimental Results and discussion

To evaluate the combinations of the preprocessing and feature extraction techniques with the classifier, a video word dataset was created as there is no standard dataset available. The video text line were extracted using our text detection technique [9]. 896 word images were extracted from the video text lines using our word segmentation method [8]. The dataset comprised of 310 Hindi, 283 Bengali and 303 English words. The results obtained using the combination of various techniques are detailed in the Tables I, II, III and IV. A five-fold cross-validation technique was used to compute the accuracy of script identification.

Various experiments were performed to find a better combination of pre-processing techniques with a feature to identify the scripts more accurately. The first experiment was performed applying the texture based features, i.e Gabor filter and Zernike moments, and gradient directional features without applying any preprocessing techniques to the word images. These experiments provide an indication of the script identification accuracy without using any pre-processing techniques and also helps to study the impact on the accuracy.

A. Performance using Gabor and Zernike features

Experiments using the Gabor filter and Zernike moment features without any pre-processing resulted in a comparatively lower accuracy as compared to the accuracy obtained using gradient directional feature. The confusion matrix using Gabor features, without any pre-processing is presented in the Table I and 55.69% accuracy was obtained using Gabor filters, which is not an encouraging result. On the other hand the accuracy obtained using Zernike moment features was around 50%, which is also bit lower than expected. The confusion matrix for Gabor features in the Table I shows that highest confusion of about 40.86% was among Bengali and Hindi. In Zernike feature, the highest confusion was also between Hindi and Bengali scripts. Upon analysis it was found that the size normalization step of both the texture based features was the main reason behind the lower accuracy. As video frames suffer from low resolution, blur, complex backgrounds, etc., size normalization sometime changes the structure of the words as well as characters and hence resulted in a lower accuracy. The substantially lower results obtained have discouraged us from employing the features on the pre-processed images.

TABLE I: Confusion matrix for script identification using Gabor feature without applying pre-processing (in %)

Scripts	English	Bengali	Hindi
English	70.85	15.61	13.54
Bengali	24.38	35.23	40.39
Hindi	12.89	26.13	60.98

B. Performance using Gradient feature

Three set of experiments were conducted using the gradient features. The first experiment (Experiment-1) was performed by extracting the gradient directional features from the original word images, without applying any pre-processing techniques. The second experiment (Experiment-2) was undertaken by applying the pre-processing techniques to the original word images and then the gradient features were extracted for classification. The third experiment (Experiment-3) was conducted to study the performance of script identification on short and long words. The experimental results obtained for each experiment are detailed below.

Experiment-1: The accuracy obtained using the gradient directional features was 86.5%. The confusion matrix obtained in the first experiment using gradient features is shown the Table II.

TABLE II: Confusion matrix for script identification using Gradient feature without applying pre-processing (in %)

Confusion matrix	English	Bengali	Hindi
English	94.14	4.21	1.65
Bengali	6.00	81.99	12.01
Hindi	2.90	13.87	83.23

Experiment-2: The second experiment was performed by applying the pre-processing techniques on the original word images and then gradient directional features were extracted. Table III presents the confusion matrix obtained by applying gradient features on both the skeleton and super-resolution images. The highest accuracy of 87.50% was obtained with the combination of super resolution technique and gradient directional features. An accuracy of 87.28% was obtained using skeleton images and gradient directional features, which is also competitive. The confidence score and accuracy analysis was also undertaken and the distribution of the script identification accuracy along with its corresponding confidence score as given by the classifier is shown in Figure 6. By confidence score, we mean to say the probability estimation of the recognized class [17]. It can be noted that about 69% of the words were classified as the correct script with a high confidence score (greater than 0.8) and only a few words (around 3%) have a lower confidence score of 0.5-0.59. Hence, gradient directional features are quite robust.

Experiment-3: The third experiment we conducted was to

TABLE III: Confusion matrix for script identification using Gradient feature (in %) on skeleton and super resolution images

Confusion Matrix	Skeleton			Super Resolution		
	English	Bengali	Hindi	English	Bengali	Hindi
English	93.95	5.50	0.55	94.52	4.22	1.26
Bengali	6.61	78.31	15.08	5.94	81.4	12.66
Hindi	1.18	9.25	89.57	2.19	11.23	86.58
Avg. Accuracy	87.28%			87.50%		

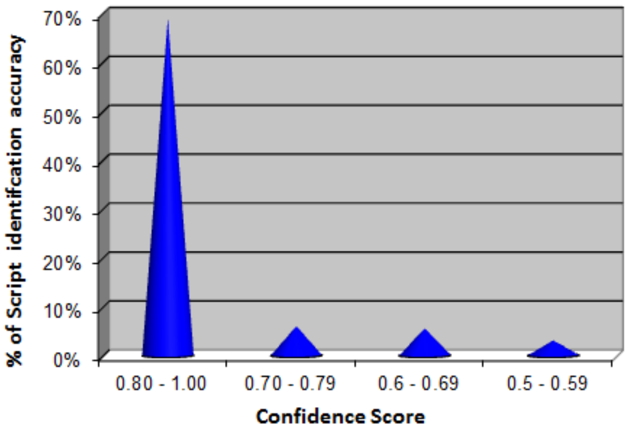

Fig. 6: Distribution of script identification accuracy with corresponding confidence score value

TABLE IV: Confusion matrix for scripts identification using Gradient feature on short and long words (in %)

Confusion Matrix	Short Words			Long Words		
	English	Bengali	Hindi	English	Bengali	Hindi
English	89.23	7.08	3.69	94.05	4.84	1.10
Bengali	5.79	76.7	17.50	7.46	84.52	7.46
Hindi	3.39	13.82	82.79	4.49	7.21	88.30
Avg. Accuracy	82.90%			89.15%		

understand the performance of script identification using gradient features on super resolution images of short words (words having three or less characters) and long words (words having more than three characters) of all the three scripts. The experimental results are shown in the Table IV. The short word dataset consists of 89, 128 and 163 words for English, Bengali and Hindi scripts, respectively. Whereas, the long word dataset consists of 214, 155 and 147 words for English, Bengali and Hindi scripts, respectively. The experimental results show that the accuracy for short word was less (82.9%), which is 4.6% less than the accuracy obtained for the combined dataset. On the other hand the accuracy for the long word was 89.15% which is nearly 1.5% more than the accuracy for the combined dataset.

The experiment showed that misclassification occurred more when there are fewer characters present in the word. As Hindi and Bengali scripts are structurally similar, they were naturally confused more amongst themselves. The inherent problem with video also contributed a bit to the misclassification. Due the presence of less structural information in the short words, the accuracy decreased substantially. The accuracy obtained using the long words increased because more script specific information is available as there are more characters in the words.

VI. Conclusion

This paper presented a study of various techniques for word-wise video script identification. A comparative study of the combination of pre-processing techniques with texture and gradient-based features was presented in the paper. SVMs were used for the classification experiments. Experiments show that the combination of the super resolution with gradient features performed better than the others and 87.50% accuracy was obtained. A large dataset with complex backgrounds was used for the experiments and the results obtained were promising. Future research plans include the usage of more Indian scripts in order to create a more robust system capable of accurately handling multiple scripts.

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] 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.

[5] 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.

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

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

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

[9] 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.

[10] A. Khotanzad and Y. H. Hong, *Invariant image recognition by Zernike moments*, IEEE Transactions on PAMI, 12(5), pp. 489-497, 1990.

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

[12] W. M. Pan, C. Y. Suen and T. D. Bui, *Script Identification Using Steerable Gabor Filters*, In Proc. ICDAR, pp. 883-887, 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] 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.

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

[16] U. Pal, N. Sharma, T. Wakabayashi, and F. Kimura, *Handwritten character recognition of popular south Indian scripts*, In Proc. of the SACH'06, LNCS 4768, pp.251-264, 2008.

[17] T.-F. Wu, C.-J. Lin, and R. C. Weng. *Probability Estimates for Multi-class Classification by Pair wise Coupling*, Journal of Machine Learning Research, 5, pp. 975-1005, 2004.

[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.