ICDAR2015 Competition on Video Script Identification (CVSI 2015)

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

***Abstract*—This paper presents the final results of the ICDAR 2015 Competition on Video Script Identification. A description and performance of the participating systems in the competition are reported. The general objective of the competition is to evaluate and benchmark the available methods on word-wise video script identification. It also provides a platform for researchers around the globe to particularly address the video script identification problem and video text recognition in general. The competition was organised around four different tasks involving various combinations of scripts comprising tri-script and multi-script scenarios. The dataset used in the competition comprised ten different scripts. In total, six systems were received from five participants over the tasks offered. This report details the competition dataset specifications, evaluation criteria, summary of the participating systems and their performance across different tasks. The systems submitted by Google Inc. were the winner of the competition for all the tasks, whereas the systems received from Huazhong University of Science and Technology (HUST) and Computer Vision Center (CVC) were very close competitors.**

Keywords*—*Video scripts identification; multi-lingual OCR; multi-lingual video-text; competition

I. Introduction

In multi-lingual and multi-script countries, the use of two or more scripts is quite common for information communication through news and advertisement videos or images transmitted across various television channels. The massive information explosion across multiple communication channels creates a very large databases of video and images. Hence, effective management of videos and images requires proper indexing for the retrieval of the relevant video and images. In general, text is usually an important constituent of the news and advertisement videos/images. Associating videos/images with the keywords found in the text appearing in them helps in their effective management and retrieval. Therefore, text present in videos/image plays an important role in automatic video indexing and retrieval. Hence, the OCR of multi-lingual video-text is crucial and a challenging task. The competition aims to find generic algorithms/system for identifying video scripts irrespective of the scripts being considered. Due to the unavailability of a universal OCR approach to recognize multi-lingual text, script identification followed by the use of appropriate script-wise OCR is a legitimate approach for recognizing the text.

The research on script identification [1], [9], [12], [10] todate 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, distortion, multiple font types and size and orientation of the text [2], [3]. Script identification from video frames has not been explored much as compared to traditional scanned documents. Recently, a few research works [4], [5], [6], [7], [8] have been published, which focus on the video script identification problem. Although there are many works on script identification [1] from scanned documents having simple backgrounds, to the best of our knowledge there is only a few of works [4], [5], [6], [13]reported in the literature on word-wise script identification from video. Samples of video frames illustrating the multi-lingual nature of videos/images are shown in Figure 1. Figure 1 demonstrates and clearly explains the necessity of script identification and the challenges involved in the OCR of text from video frames/images.

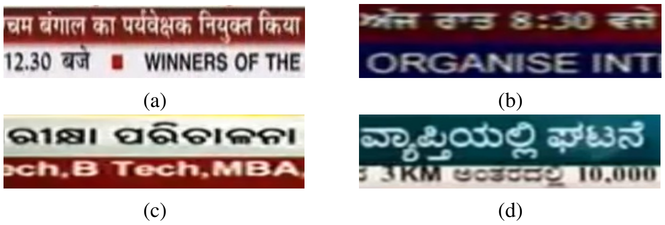


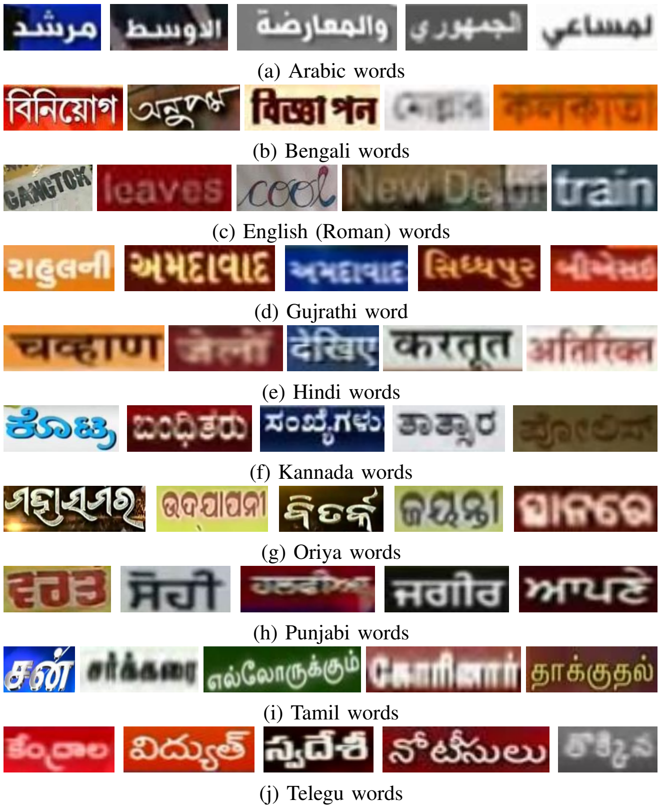
Fig. 1: Samples of video frames showing the multi-lingual nature

The Competition on Video Script Identification (CVSI 2015) makes the following contributions to the video/camera-based document analysis community:

- A new multi-lingual video word dataset comprising ten scripts
- Provides a common platform for system evaluation
- Benchmarking existing and participating systems
- The facilitation of research in video script identification and text recognition in general.

The competition is organized around four tasks: tri-script identification (Task-1), North Indian script identification (Task-2), South Indian script identification (Task-3) and multi-script identification (Task-4).

This report details the performances of the systems received for evaluation. The organization of the rest of this paper is as



(a) Arabic words

(b) Bengali words

(c) English (Roman) words

(d) Gujrathi word

(e) Hindi words

(f) Kannada words

(g) Oriya words

(h) Punjabi words

(i) Tamil words

(j) Telegu words

Fig. 2: Samples of video word images taken from the multi-lingual video word competition dataset

follows: In Section II, a description of the dataset is provided. The tasks involved in the competition are presented in Section III. Section IV presents a performance evaluation protocol. Section V discusses the systems submitted by the participants. The outcome in terms of performance of the submitted systems are described and analysed in Section VI. Finally, conclusions are drawn in Section VII.

II. Competition Dataset

On reviewing the existing work on video script identification, it was identified that there is no publicly available dataset on multi-lingual video words. Hence, a multi-lingual video word dataset was created and published for benchmarking the existing systems available for video script identification. Ten scripts namely, English (Roman), Hindi (Devnagari), Bengali, Oriya, Punjabi (Gurumukhi), Gujrathi, Arabic, Kannada, Tamil and Telegu were considered for creating the dataset. The video words [11] were extracted from the multi-lingual video text lines and the script ground truth was generated manually. The dataset provided to the participants for the experiments consists of 10688 word samples from the ten scripts. Statistics of the dataset are given in Table I. The dataset was divided into training (60%), validation (10%) and testing (30%) dataset, randomly. The training dataset comprises 6412 samples, the validation set of 1069 samples, and the test dataset with 3207 samples. The dataset was published as a part of the ICDAR 2015 competition on Video Script Identification (CVSI-2015) [15]. The test dataset is a closed dataset and is made available once the participants submit their systems for evaluation.

TABLE I: Multi-Script Video Word Dataset Statistics

Script	Total word	Training set	Validation set	Testing set
Arabic	1011	607	101	303
Bengali	1032	619	103	310
English (Roman)	1135	681	113	341
Gujrathi	1086	651	108	327
Hindi (Devnagari)	1088	653	109	326
Kannada	1047	628	105	314
Oriya	1087	652	109	326
Punjabi (Gurumukhi)	1055	633	106	316
Tamil	1070	642	107	321
Telegu	1077	646	108	323
Total	10688	6412	1069	3207

A few samples of video words taken from the dataset for each script are shown in Figure 2. Figure 2 also shows that the text appearing in the word images are in different orientations, fonts, size and suffer from low resolution and blur. This adds more complexity to script identification and text recognition from videos.

III. Competition Tasks

An Indian state generally uses three official languages. For example, the West Bengal State of India uses Bangla, Hindi and English as official languages. Hence, a single document may contain one or more of these three scripts. Four different tasks were offered to the participants. Task-1 is for identifying a script from script triplets. In Task-2 and Task-3, we have divided all of the scripts into two classes based on their regions of use namely north and south Indian scripts. Finally we have considered all of the scripts in Task-4. A brief description of the tasks are given below.

Task-1: Identifying scripts from eight different script triplets (Combinations of three scripts, keeping English and Hindi in all the combinations), based on their use in the Indian sub-continent.

Task-2: Identifying the combination of scripts used in north India. This task involves identification of seven scripts, namely, English, Hindi, Bengali, Oriya, Gujrathi, Punjabi and Arabic.

Task-3: Identifying the combination of scripts used in south India. This task involves identification of five scripts, namely, English (Roman), Hindi, Kannada, Tamil and Telegu.

Task-4: Identifying scripts from the combination of all the ten scripts is the challenge in Task-4. Three south Indian scripts (i.e. Kannada, Tamil and Telugu) and six north Indian scripts (i.e. Hindi, Bengali, Oriya, Gujrathi, Punjabi and Arabic) with English scripts have been considered for Task-4 of the video script identification competition.

IV. Evaluation Procedure

Participants were allowed to take part in single or multiple tasks out of the four tasks arranged for the competition. Each task is evaluated and ranked separately and the performance evaluation and system ranking are based on individual tasks. The identification accuracies in terms of percentages are the decisive factors of the competition winner. The system which performs best for individual tasks is declared the winner for that particular task. The system performance accuracy has been calculated as follows

$$Accuracy = (CC/GT) \times 100$$

Where CC is the number of correctly classified sample from the test dataset; GT represents the ground truth as well as the total number of samples in the test dataset. The participants have submitted their systems / executables either in Windows (XP or Win 7) executable format or Linux/Unix. The resulting text files have been created by the submitted systems for each participated task which contains the following format and saves the identification result for all of the samples in the test dataset:

$$[TestSampleName]|[IdentifiedScript]$$

where TestSampleName represents the test file name with the path information and the system returns the first three characters of the identified scripts (e.g. $Test1.jpg|Eng$ or $Test2.jpg|Ben$, etc). The three characters of identified scripts can be anyone in the set $\{Arb, Ben, Eng, Guj, Hin, Kan, Ori, Pun, Tam, Tel\}$ for the respective tasks.

V. Participating Systems

Six systems were received for evaluation from five participants (one of the participant from CVC, Spain submitted systems using two different algorithms) to the ICDAR 2015 video script identification competition. Manjunath Shantharamu from Central University of Kerala (CUK) participated in three tasks only (Task-2, Task-3 and Task-4) and the rest of the participants participated in all of the four tasks. Affiliations of the participants and brief descriptions of their systems are presented as follows.

1) **C-DAC, India:** Swapnil Belhe from Centre for Development of Advanced Computing (C-DAC), has participated in all of the four competition tasks. The submitted systems convert images into gray scale. Next, two different features namely, Histogram of Oriented Gradients (HoG) and Linear Binary Pattern (LBP) are computed from the video script sample images. These two features are finally combined to obtain the 292 (36 HoG features and 256 LBP features) dimensional feature used for training and testing purposes. A SVM classifier with s Radial Basis function kernel has been employed for the task of multi-class classification.
2) **HUST, China:** Baoguang Shi, and his team members Cong Yao, Chengquan Zhang, Wei Shen, Zheng Zhang, and Xiang Bai from Huazhong University of Science and Technology (HUST) have submitted systems by participating in all of the four competition tasks. The system's algorithm is mainly based on a deep neural network which is a variant of the convolutional neural network. The system takes input images of arbitrary aspect ratios and can precisely predict script types from text images.
3) **CVC-1, Spain:** Lluis Gomez from the Computer Vision Center at UAB have participated in all of the four competition tasks and submitted two sets of systems for video script identification. The first set of systems used an algorithm named CVC_DAG_UFL_NBNN, and single layer Convolutional Neural Network has been trained for script identification and a Naive Bayes Nearest Neighbor (NBNN)-based classification technique is employed.
4) **CVC-2, Spain:** The second set of systems from CVC, submitted by the same participant named CVC_DAG_UFL_I2CMLNBNN were almost same as the first algorithm but Mahalanobis distances using the Large Margin Metric Learning algorithm were employed instead of Euclidean distance in the Nearest Neighbor classifier.
5) **Google, Inc.:** Yuanpeng Li from Google has participated in all four of the competition tasks. In their system, an input image is scaled to a fixed height and binarized, then passed to a deep convolutional network for class prediction. If an image is sufficiently wide, a sliding window is used and the class having the highest confidence is chosen. During training, Stochastic Gradient Descent and L2 regularization are used, and the training data is augmented by introducing replicas at various resolutions, widths, and degrees of stroke weight.
6) **CUK, India:** Manjunath Shantharamu from Central University of Kerala (CUK) has participated in three tasks namely Task-2, Task-3 and Task-4. In the submitted systems, a video word image was pre-processed using a K-means clustering-based technique to identify text and non-text regions. The features are clustered into two clusters to segment the video frame into foreground and background components. HoG features [14]were extracted from the text regions and the Nearest Neighbour classifier was used for video script identification.

VI. Results and Discussion

The performance of the systems from all of the participants are evaluated and reported in this section. As mentioned in Section V, four participants have participated in Task-1 and five participants have participated in the rest of the tasks (Task-2, Task-3 and Task-4) given for the video script identification competition. Eight systems for the eight different script triplets were submitted by each of the four participants for Task-1 of the competition. Three different systems have been submitted by each of the five participants for Task-2, Task-3 and Task-4 of the competition. There is an extra set of systems for each of the tasks from CVC-2 as mentioned in Section V.

1. Performance on Task-1: Eight different script triplets were given for identification in Task-1 of the competition. The results obtained from all of the participating systems for Task-1 are presented in Table II. The systems submitted by Google Inc. have outperformed other systems in the six different script triplets, except the triplet combination comprising Kannada (Com4) and Oriya (Com5). The best performance obtained for each of the script triplets is highlighted with boldface font in Table II. The systems submitted by the HUST have outperformed other participating systems in the two script

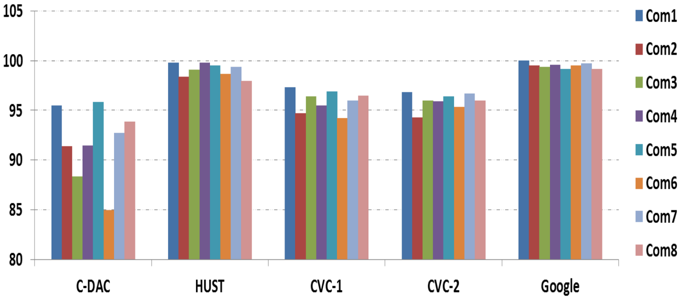


Fig. 3: Graphical representation of the performance of participating systems on tri-script combinations

triplets having Kannada and Oriya. The lowest overall accuracy of 99.19% was obtained in Task-1 by the Google Inc. system for the identification of Telugu script triplets. Table II shows that systems from Google Inc. and HUST have a very competitive performance. The lowest performance for all the script triplet combinations was obtained for the C-DAC's system. A graphical representation of the performance of all the systems in Task-1 are shown in Fig. 3.

TABLE II: Results of Task-1 (Video script identification from eight different script triplets). Here English (Roman) and Hindi are common in all of the script triplets.

		Accuracy (%)				
Script triplets		C-DAC	HUST	CVC-1	CVC-2	Google
Com1 Samples:970	Arabic	98.68	**100**	99.34	99.67	**100**
	English	92.67	99.41	93.55	91.50	**100**
	Hindi	95.4	**100**	99.39	99.69	**100**
	Overall	95.46	99.79	97.32	96.8	**100**
Com2 Samples:977	Bengali	98.71	97.1	96.13	93.54	**99.68**
	English	89.44	99.12	90.62	90.62	**99.70**
	Hindi	86.50	98.77	97.55	98.77	**99.08**
	Overall	91.40	98.36	94.68	94.27	**99.49**
Com3 Samples:994	Gujrathi	92.05	99.39	98.47	98.78	**99.69**
	English	78.00	97.95	92.96	90.62	**98.53**
	Hindi	95.40	**100**	97.85	98.77	**100**
	Overall	88.33	99.09	96.38	95.98	**99.4**
Com4 Samples:981	Kannada	88.54	99.36	99.04	97.77	**99.68**
	English	89.74	**100**	90.91	90.91	99.41
	Hindi	96.01	**100**	96.93	99.39	99.69
	Overall	91.44	**99.80**	95.51	95.92	99.59
Com5 Samples:993	Oriya	98.47	98.77	**99.39**	98.47	98.47
	English	92.38	**99.71**	92.08	91.20	99.12
	Hindi	96.93	**100**	99.39	99.69	**100**
	Overall	95.87	**99.50**	96.88	96.37	99.19
Com6 Samples:983	Punjabi	90.82	98.10	98.10	97.15	**99.68**
	English	93.25	**99.41**	92.67	91.49	99.12
	Hindi	70.55	98.47	92.02	97.54	**99.69**
	Overall	84.94	98.68	94.2	95.32	**99.49**
Com7 Samples:988	Tamil	98.44	**100**	bf 100	**100**	**99.69**
	English	83.87	98.24	90.91	90.91	**99.41**
	Hindi	96.32	**100**	97.23	99.39	**100**
	Overall	92.71	99.39	95.95	96.66	**99.70**
Com8 Samples:990	Telugu	98.76	99.07	**99.69**	98.14	99.38
	English	87.98	95.01	90.62	90.32	**98.83**
	Hindi	95.09	**100**	99.38	99.69	99.39
	Overall	93.84	97.98	96.46	95.96	**99.19**

2. Performance on Task-2: Six systems were submitted by the five participants for Task-2 on video script identification from six North Indian scripts along with English script. The evaluated performance of all the submitted systems are presented in Table III. Script-wise system's accuracy along with the overall accuracy is presented for all of the scripts considered in the task. The system submitted by Google Inc. has outperformed other systems with 99.19% accuracy and the script-wise best accuracy is highlighted in boldface font. Systems submitted by HUST, CVC-1 and Google Inc. have identical performance for Oriya script, which is 99.39%. For Arabic, 100% accuracy was achieved by both Google Inc's and HUST's systems, whereas, 98.78% accuracy was obtained for Gujrathi by Google Inc. and CVC-2. A graphical representation of the performance of all the systems on Task-2 are shown in Fig. 4.

TABLE III: Results of Task-2 (Video script identification from six North Indian scripts along with English (Roman) script.

	Accuracy (%)					
Scripts	**C-DAC**	**HUST**	**CVC-1**	**CVC-2**	**Google**	**CUK**
Arabic	99.34	**100.00**	99.67	99.67	**100**	95.71
Bengali	93.23	96.45	94.52	92.58	**99.68**	69.03
English	75.95	97.07	90.03	89.74	**99.41**	74.49
Gujrathi	88.99	97.86	98.17	**98.78**	**98.78**	80.73
Hindi	69.93	96.32	91.41	96.32	**98.16**	62.88
Oriya	91.72	**99.39**	**99.39**	98.16	**99.39**	82.52
Punjabi	90.19	96.84	97.47	96.52	**99.05**	92.41
Overall	86.79	97.69	95.73	95.91	**99.19**	79.50

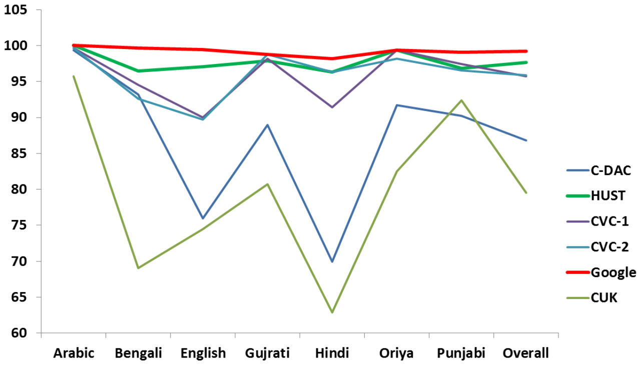


Fig. 4: Graphical representation of the performance of participating systems on North Indian scripts

3. Performance on Task-3: Three South Indian scripts along with English and Hindi scripts were considered for Task-3 of video script identification. Table IV shows the performance obtained from the six systems submitted by all of the five participants. The best overall accuracy of 98.95% was obtained by the Google's system for South Indian script identification. The second best accuracy (97.53%) has been achieved by HUST's system. For Tamil script, 100.00% accuracy was achieved by the HUST, CVC-1, and CVC-2 groups. In contrast, HUST and Google achieved 100.00% for Hindi script. In this task Google and HUST were very close competitors. Fig. 5 shows the graphical representation of the performance of all the submitted systems on Task-3.

4. Performance on Task-4: All of the ten scripts (North and South Indian) are considered for Task-4 on video script

TABLE IV: Results of Task-3 (Video script identification on South Indian scripts along with English (Roman) and Hindi scripts).

	Accuracy (%)					
Scripts	**C-DAC**	**HUST**	**CVC-1**	**CVC-2**	**Google**	**CUK**
English	76.54	95.01	88.56	89.44	**99.12**	71.84
Hindi	96.93	**100**	96.01	99.07	**100.00**	93.25
Kannada	69.75	93.95	**97.45**	96.82	96.82	75.48
Tamil	95.64	**100.00**	**100.00**	**100.00**	99.07	91.59
Telugu	95.98	98.76	95.36	93.81	**99.69**	63.78
Overall	86.95	97.53	95.38	95.75	**98.95**	79.14

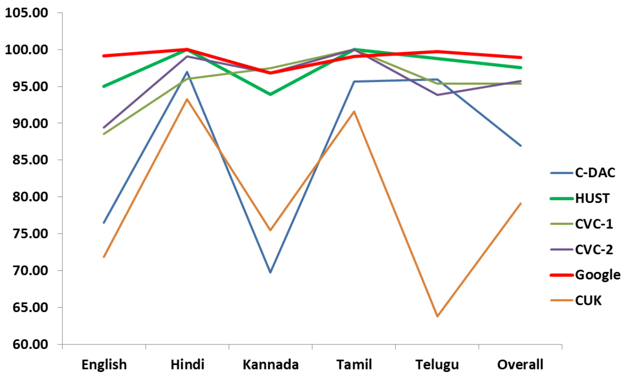


Fig. 5: Graphical representation of the performance of participating systems on South Indian script

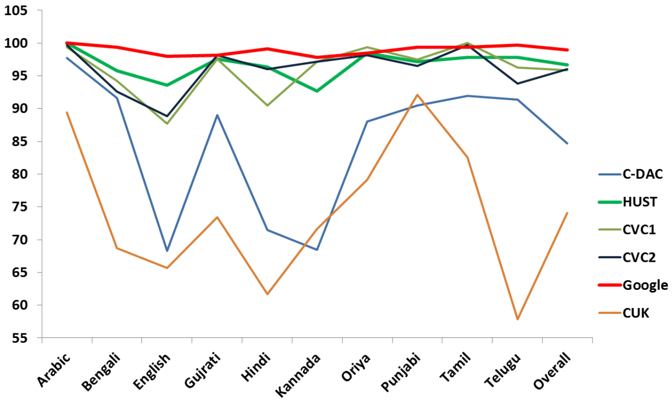


Fig. 6: Graphical representation of the performance of participating systems on all Indian scripts

identification. Table V shows the performance obtained from the systems submitted by all of the five participants. The best overall accuracy of 98.91% was obtained by Google's system for Task-4 of video script identification. The second best accuracy i.e. 96.69% was achieved by HUST's system. Additionally, it can be noted that 100% accuracy was achieved by HUST and Google's systems for Arabic scripts. In contrast, 100.00% accuracy was achieved by CVC-1 for Tamil script. In this task also, HUST and Google had very close results. Figure 6 shows the graphical representation of the performance of all the submitted systems on Task-4.

TABLE V: Results of Task-4 (Video script identification considering all the ten scripts)

	Accuracy (%)					
Scripts	**C-DAC**	**HUST**	**CVC-1**	**CVC-2**	**Google**	**CUK**
Arabic	97.69	**100.00**	99.34	99.67	**100.00**	89.44
Bengali	91.61	95.81	94.19	92.58	**99.35**	68.71
English	68.33	93.55	87.68	88.86	**97.95**	65.69
Gujrathi	88.99	97.55	97.55	**98.17**	**98.17**	73.39
Hindi	71.47	96.31	90.49	96.01	**99.08**	61.66
Kannada	68.47	92.68	97.13	97.13	**97.77**	71.66
Oriya	88.04	98.47	**99.39**	98.16	98.47	79.14
Punjabi	90.51	97.15	97.47	96.52	**99.38**	92.09
Tamil	91.90	97.82	**100.00**	99.69	99.38	82.55
Telugu	91.33	97.83	96.28	93.80	**99.69**	57.89
Overall	84.66	96.69	95.88	96.00	**98.91**	74.06

VII. Conclusions

Five groups took part in the ICDAR 2015 competition on video script identification which was organised to evaluate the existing and recently proposed methods on video script identification. In this paper, the performance of all the submitted systems are reported on the tasks given to the participants. The best performance was achieved by Google's systems submitted by Yuanpeng Li from Google, Inc. for all of the competition tasks. We plan to make the competition dataset even larger and more challenging for future competitions and look forward to more participants and interest in this exciting research area.

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, U.Pal, M. Blumenstein, *A Study on Word-Level Multi-script Identification from Video Frames*, In Proc. IJCNN, pp. 1827-1833, 2014.

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

[6] P. Shivakumara, N. Sharma, U. Pal, M. Blumenstein, and C. L. Tan, *Gradient-Angular-Features for Word-wise Video Script Identification*, In Proc. ICPR, pp. 3098-3103, 2014.

[7] 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.

[8] 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.

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

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

[11] 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.

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

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

[15] http://www.ict.griffith.edu.au/cvsi2015/