StreetOCRCorrect: An Interactive Framework for OCR Corrections in Chaotic Indian Street Videos

Pankaj Singh[1] *IIT Bombay* Mumbai, India
Bhavya Patwa[2] *Ahmedabad University* Ahmadabad, India
Rohit Saluja[3] *IITB-Monash Research Academy* Mumbai, India
Ganesh Ramakrishnan[3] *IIT Bombay* Mumbai, India
Parag Chaudhuri[3] *IIT Bombay* Mumbai, India

[1] pr.pankajsingh@gmail.com, [2]bhavya.p.btechi14@ahduni.edu.in, [3]{rohitsaluja, ganesh, paragc}@cse.iitb.ac.in

***Abstract*—Obtaining a high-quality OCR output in smart cities, with human-in-the-loop, is an interesting problem for surveillance and other similar applications. Achieving high accuracy while reading license plates in the real world videos is cumbersome due to complexities like multiple vehicles, high-density traffic in spatial and temporal domains, varying camera angles and illumination, occlusions and multiple resolutions. We present a modular framework for OCR corrections in the chaotic Indian traffic videos that especially involve complex license plate patterns. Such patterns are obtained from a state-of-the-art deep learning model trained on video frames. Since such a model reads the text from videos (instead of images), we incorporate multi-frame consensus for generating suggestions in our framework. To ease the correction process, our human-interactive framework first breaks down the multi-vehicle videos into multiple clips, each containing a single vehicle from the video using an object detector and a tracker. Our framework then provides suggestions for an individual vehicle using multi-frame consensus. Our framework then selectively presents these extracted clips to the user to verify/correct the predictions with minimal human efforts via interactive suggestions. Such high-quality output can be used to continuously update a large database for surveillance and can be further used to improve the accuracy of deep models in the complex real-world scenarios.**

I. Introduction

Reading text in images is a challenging research problem due to complexities like varying camera angles and illumination, font size and type, varying backgrounds, occlusions and multiple resolutions. Early works on scene text recognition invoke text detection and recognition in separate phases [1], [2], [3], [4], [5], [6]. Recent text spotters include deep models that are trained end-to-end, but with supervision at the level of text as well as at the level of words and text-boxes [4], [7]. The two recent end-to-end approaches in this direction, that directly work on the complete scene image, without supervision at level of text boxes, are:

1) STN-OCR: A single Neural Network for Text Detection and Text Recognition [8]: The work presents a model that extract the grids with the location of word images via a spatial transformer network. The grid is then applied to the complete image to obtain the word images and the cropped word images are passed to another spatial transformer network for recognition. This work does not need supervision at the level of detection.
2) Attention-OCR [9]: This work is based on inception based encoder and attention based Recurrent Neural Network (RNN) decoder. This work is interesting since it does not involve any cropping of word images. The model performs character-level recognition directly on the complete scene image thus utilizing the global context while reading the scene. This model is an open source tensorflow (a popular library for deep learning) implementation. There are some variants available based on how the attention mechanism is being used. OCR-on-the-go is one such variant which uses multi-head attention to enhance the accuracy on license plates as

Fig. 1. Sample frames from chaotic street videos with variations like multiple vehicles, high density and occlusions, with multi-line/variable-sized or cursive/hand-written characters, darkness, non-rectangular/old-dusty plates.

978-1-7281-5054-3/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDARW.2019.10036

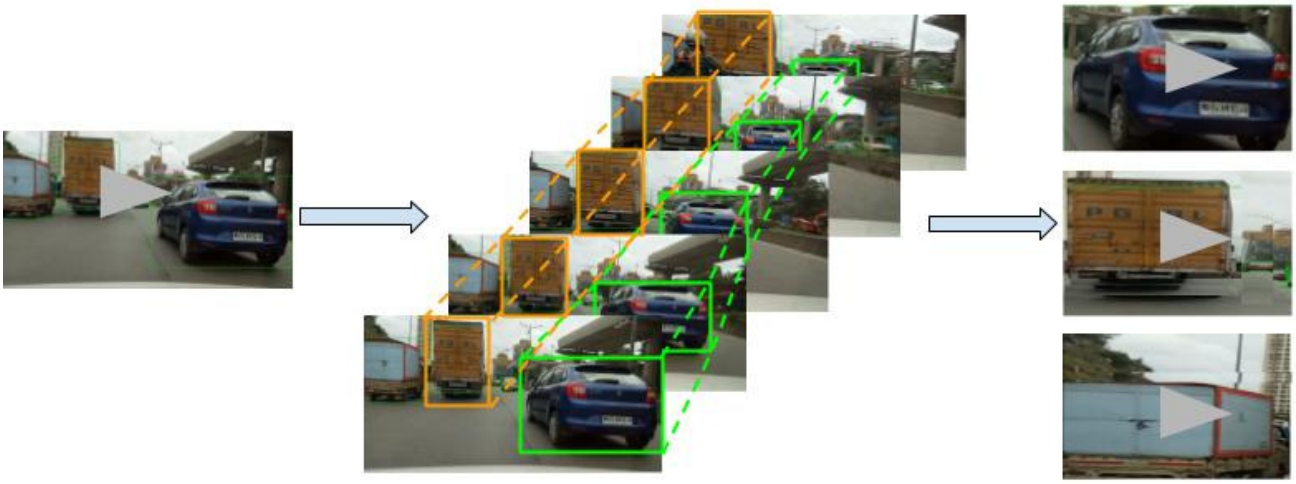

Fig. 2. Breakdown of a video by our framework in (i) spatial domain as shown in two (for illustration) different colors, and (i) temporal domain as shown in dotted curves. The output of this module is a dataset of multiple clips, each containing a unique vehicle.

well as street board signs [10].

These works consist of experiments on a French Street Name Signs (FSNS) dataset, for which the later work, outperforms all other methods. Any of such end-to-end models are well suited for the backend of our framework.

We target to address a more specific scene text recognition problem in the Indian context, viz., correcting the license plates text in traffic videos. Automatic License Plate Recognition (ALPR) is an important component for smart cities. Various smart city applications such as automated parking, law and order enforcement, surveillance, access management, toll tax collection, intelligent transport system and traffic flow management require high quality ALPR systems.

The problem is especially challenging in the Indian context owing to conditions such as chaotic traffic, multi-line license/variable sized characters, cursive/handwritten fonts and non-rectangular/old-dusty license plates. Several of these are illustrated in the Figure 1. It's even more difficult to recognize plates in the videos due to motion blur, varying orientations and low-resolution cameras. License plate recognition is generally performed in two steps: 1) License plate detection, and 2) License plate recognition. Features such as edge, texture, color, etc. are generally used for license plate detection [11]. However, features based approaches suffer from the problem of an excess of false positives. Character level recognition based on neighbourhood connectivity, projections, template matching, etc. were used in early work for license plate recognition. However, these methods suffer from the problem of segmentation error due to overlapping characters and confusions owing to visually similar n-grams. Thus the literature moved gradually from character based to sliding window based and then finally to word level recognition via Neural Networks. In a recent work on Indian License Plate recognition, edge based features have been used [12]. The work then uses a Convolutional Neural Network (CNN) to discard false positives. The license plates are then cropped from the scenes and passed to another CNN for recognition.

OCR text correction has been the theme of many works [13], [14]. Competitions on historical documents in a total of ten European languages further highlight the importance of PostOCR Work [15], [16]. There have been tools and works on OCR correction, specifically related to Indian documents. [17], [18], [19]. Recent work on developing a tool for annotating a large amount of scene text images discuss the rareness of such tools [20]. We, however, further develop a tool for annotating a large number of videos (as well as images) for license plate text. Various video annotations tools are available however either they work for predefined set of classes (i.e. don't support text annotations) or can not provide annotation at sub-frame (object) level [21], [22].

Obtaining large scale multi-frame video annotations is a challenging problem due to unreliable OCR systems as well as expensive human efforts. The predictions obtained on videos by most OCR systems are fluctuating as can be observed in https://youtu.be/VcNSQGO0j7s and https://youtu.be/LXGXOGryuc4 [4]. The fluctuations can also be present due to various external factors such as partial occlusions, motion blur, complex font types, distant text in the videos. Thus, such OCR outputs are not reliable for downstream applications such as surveillance, traffic law enforcement and cross border security system. We present our framework for correcting complex license plate patterns (refer Figure 1) in street videos, that we call as StreetOCRCorrect. The high-quality output obtained from such a framework can be used to prepare a large database. Such a database can enable new applications like reliable text based and image (as well as text) based search systems, analytic dashboard, traffic flow monitoring, etc. Such a database can also be used to continuously improve the OCR models.

We present the description of our framework in Section II.

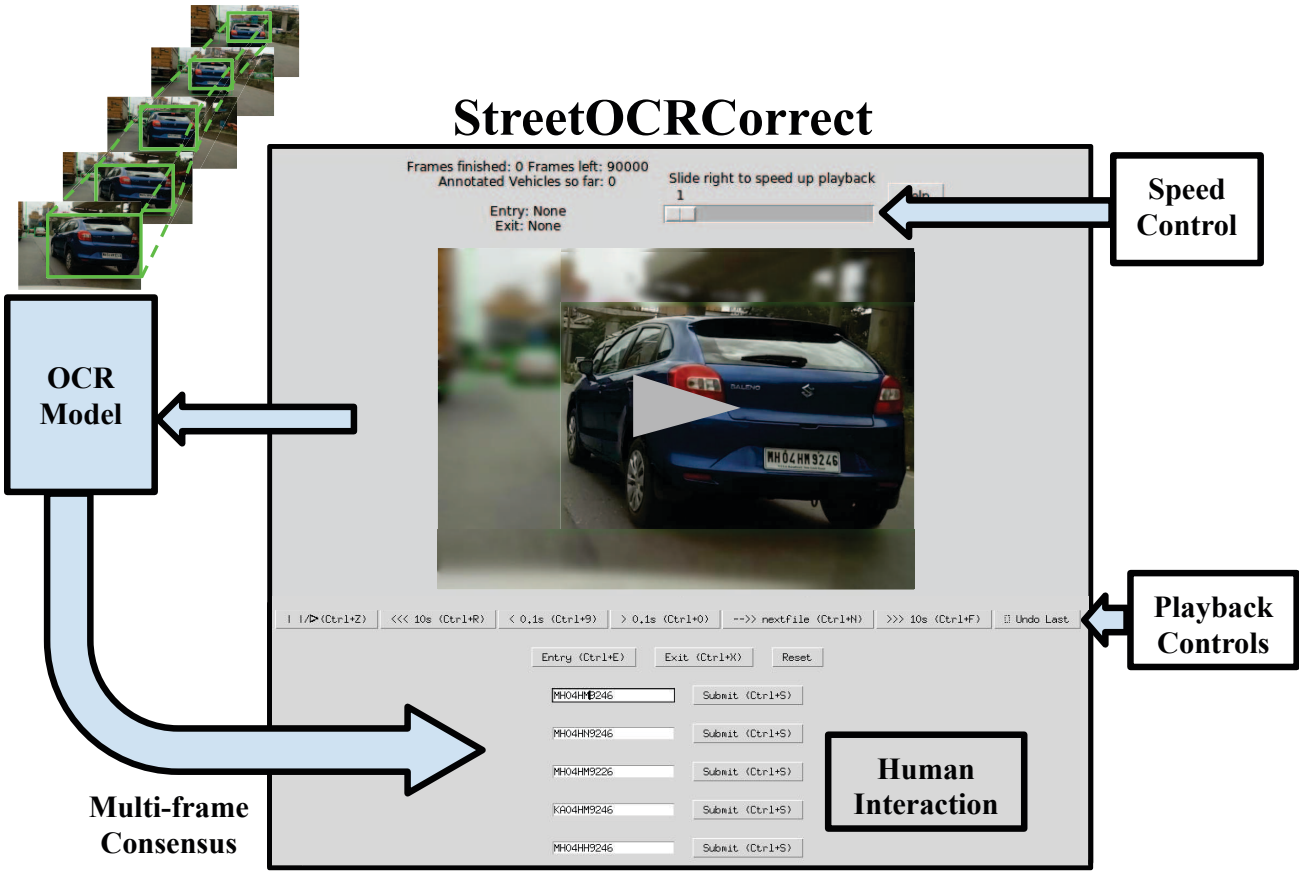

Fig. 3. Components of our framework.

Sample results from our model are presented in Section III. We then conclude the paper with discussion on future work in Section IV.

Some of our key contributions are as follows:

1) We present StreetOCRCorrect: a novel framework for OCR corrections in chaotic street videos. Our modular framework is available on github at https://github.com/rohitsaluja22/StreetOCRCorrect.
2) We simplify the task of correcting multiple predictions of a vehicle in videos.
3) We provide various control panels for speed control, playback control and human interaction in our framework.
4) We present the sample results on multi-frame consensus to demonstrate the use case.

II. Framework Description

Robust systems often fail in chaotic Indian scenes due to the presence of extreme variations in spatial and temporal domains like high vehicle density, multiple frames, irregular traffic conditions, occlusions etc. We present a modular framework to handle such variations.

The first module of our framework breaks down the video in a spatio-temporal domain. The spatial breakdown is achieved by using the You Only Look Once (YOLO) object detector [23]. Our framework further breaks down the video in the temporal domain using MedianFlow video tracker and creates multiple clips each containing a single vehicle within a fixed window. This is illustrated in Figure 2 The partial-occlusions are handled by blurring the surrounding of the box that contains the vehicle under consideration as shown in Figure 4. The first module removes unnecessary complexities related to handling multiple vehicles and hence enables the user to focus on correcting the license plate text for a single vehicle as shown Figure 3.

The second module of our framework enables the user to verify/correct the predictions of the extracted clips with minimal human efforts via interactive suggestions. These suggestions are obtained by using multi-frame consensus on the output obtained by applying OCR model on the clips

GT - KA01MJ3793
KA01MJ3793 - 47
KA01AJ3793 - 06
(a) KA01NJ3793 - 01

GT - MH03CW4026
MH03CW4026 - 16
MH03CV4026 - 04
(b) MH03GVW4028 - 02

GT - MP04GA8692
MP04GA8692 - 06
MP04GH692 - 03
(c) MP04GA6527 - 02

GT - MH47A4487
MH47A4487 - 66
MH47A4481 - 05
(d) MH47A4482 - 04

Fig. 4. Sample inputs, extracted from chaotic scenes, given to our framework. The top line in each caption is ground truth. The next three lines include the top 3 predictions/suggestions, along with their counts, among continuous sample video frames.

generated in the first module. We can further reduce the human intervention by selectively presenting the clips with low confidence i.e. consensus scores. As shown in Figure 3, our framework contains a speed control panel, a playback control panel and a human interaction panel for improving the user experience. Here are the key features of our framework:-

1) As the user opens the framework, and loads the folder (through config file), the video of a vehicle is shown to the user, along with the suggestions from our model in "Human Interaction" area.
2) If the license plate is visible and readable throughout the video, the user can directly verify or correct one of the suggestion and submit the result.
3) If the license plate is not visible or readable at starting or end of video, the user select the good frames by entry and exit buttons, and then verify or correct one of the suggestion and submit the result.
4) The user then clicks on the next video to upload the video of next vehichle and repeat the above steps till all the vehicles are annotated.

As a part of our framework, the third module stores the verified/corrected text with other metadata such as timestamp, GPS location (if available), etc.

The source code of our framework can be downloaded from https://github.com/rohitsaluja22/StreetOCRCorrect. We also share the links of 22 youtube videos with chaotic Indian traffic scenes, along with our code. These videos can be further used to improve the ALPR dataset complexity and frameworks like ours.

III. Video Results

The demo video on our gitHub link https://github.com/rohitsaluja22/StreetOCRCorrect shows the performance of our backend model on various challenging scenarios. Since we use the end-to-end model for recognition, we pass the vehicle image in the box with the additional region on the left and the right (blurred to avoid other license plates) through our model. The additional region is included to maintain the fixed size as well as aspect ratio in frames of the clipped videos. We detect vehicles on each frame of the videos as explained in Section II. This is shown in Figure 4. In the figure, we also present the top three predictions, along with their count, for some of the continuous sample frames for different vehicles. As shown, the predictions of our model are correct for the majority of frames on these samples. We use this system to generate license plate dataset which we further use to improve the recognition accuracy of deep learning based license plate recognition model used in OCR-on-the-go [10]. The model was trained on the video data obtained from 3 different sources at 3 different weather conditions. Since the deep learning techniques require large amount of data, we additionally collected 100 hours of traffic video data from 15 different sources. With 5 annotators working for total of 15 hours each and 3 reviewers working for 8 hours each, we generate 2.67 million high quality image level labelled dataset. Thus we generate the large amount of dataset in less than 100 man-hours, 80% of which we used for domain adaptation of the pre-trained OCR-on-the-go model. The dataset helps us in improving the sequence accuracy (exact match) of attention-ocr model from 41% to 81% on the testset (20 hours of video). We also observe that it takes 4 hrs to manually annotate the sample 1 hour video whereas we annotate the same 1 hour video in 55 minutes using the StreetOCRCorrect. This demonstrates the effectiveness of multi-frame consensus used in our framework.

IV. Conclusions

We designed an interactive framework for large scale OCR corrections in Indian Street Videos. Our framework leverages the state-of-the-art detector and tracker to ease the correction process. We further use multi-frame consensus to detect errors and reduce the cognitive load significantly via suggestions. Our framework further maintains a large scale database of high quality text which can be used in various downstream applications. We use the gold quality dataset obtained via StreetOCRCorrect to improve the sequence accuracy of the ALPR model by 40% on a complex test set containing video frames obtained from 15 different sources. As a future work, we would like to extend this work to street board signs as well as general scene text recognition. We would also like to leverage error detection models to actively improve the models.

V. Acknowledgement

We thank Mayur Punjabi for improving the UI.

REFERENCES

[1] M. Mathew, M. Jain, and C. Jawahar, "Benchmarking scene text recognition in devanagari, telugu and malayalam," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 7. IEEE, 2017.

[2] A. R. Chowdhury, U. Bhattacharya, and S. K. Parui, "Text detection of two major indian scripts in natural scene images," in *International Workshop on Camera-Based Document Analysis and Recognition*. Springer, 2011, pp. 42–57.

[3] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu *et al.*, "Icdar 2015 competition on robust reading," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 2015.

[4] M. Bušta, L. Neumann, and J. Matas, "Deep textspotter: An end-to-end trainable scene text localization and recognition framework," *International Conference on Computer Vision*, 2017.

[5] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, 2017.

[6] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou, "Focusing attention: Towards accurate text recognition in natural images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5076–5084.

[7] M. Bušta, Y. Patel, and J. Matas, "E2e-mlt-an unconstrained end-to-end method for multi-language scene text," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 127–143.

[8] C. Bartz, H. Yang, and C. Meinel, "Stn-ocr: A single neural network for text detection and text recognition," *arXiv preprint arXiv:1707.08831*, 2017.

[9] Tensorflow, "Attention Ocr Model," https://bit.ly/2BczGN3. Last accessed on March 7, 2019.

[10] R. Saluja, A. Maheshwari, G. Ramakrishnan, P. Chaudhuri, and M. Carman, "OCR On-the-Go: Robust End-to-end Systems for Reading License Plates & Street Signs," in *15th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2019.

[11] S. Du, M. Ibrahim, M. Shehata, and W. Badawy, "Automatic license plate recognition (alpr): A state-of-the-art review," *IEEE Transactions on circuits and systems for video technology*, vol. 23, no. 2, 2013.

[12] V. Jain, Z. Sasindran, A. Rajagopal, S. Biswas, H. S. Bharadwaj, and K. R. Ramakrishnan, "Deep automatic license plate recognition system," in *Tenth Indian Conference on Computer Vision(ICVGIP)*. ACM, 2016.

[13] I. Kissos and N. Dershowitz, "OCR Error Correction Using Character Correction and Feature-based Word Classification," in *12th IAPR Workshop on Document Analysis Systems (DAS)*, 2016, pp. 198–203.

[14] J. Evershed and K. Fitch, "Correcting Noisy OCR: Context Beats Confusion," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 2014, pp. 45–51.

[15] G. Chiron, A. Doucet, M. Coustaty, and J.-P. Moreux, "Icdar2017 competition on post-ocr text correction," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1. IEEE, 2017, pp. 1423–1428.

[16] ICDAR, "Competition on Post-OCR Text Correction," https://sites.google.com/view/icdar2019-postcorrectionocr. Last accessed on March 7, 2019.

[17] R. Saluja, D. Adiga, G. Ramakrishnan, P. Chaudhuri, and M. Carman, "A framework for document specific error detection and corrections in indic ocr," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 4. IEEE, 2017, pp. 25–30.

[18] V. Vinitha and C. Jawahar, "Error Detection in Indic OCRs," in *12th IAPR Workshop on Document Analysis Systems (DAS)*, 2016.

[19] R. Saluja, D. Adiga, P. Chaudhuri, G. Ramakrishnan, and M. Carman, "Error detection and corrections in Indic OCR using LSTMs," *International Conference on Document Analysis and Recognition (ICDAR)*, 2017.

[20] L. Lenc, J. Martínek, and P. Král, "Tools for semi-automatic preparation of training data for ocr," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2019, pp. 351–361.

[21] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *International Journal of Computer Vision*, pp. 1–21, 10.1007/s11263-012-0564-1. [Online]. Available: http://dx.doi.org/10.1007/s11263-012-0564-1

[22] M. Kipp, "Anvil-a generic annotation tool for multimodal dialogue," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.