MULTILINGUAL SUBTITLE TEXT DETECTION AND REMOVAL USING VIDEO INPAINTING

Department of Computer Science and Engineering
PES University, RR Campus, Bengaluru - 560085

PROBLEM STATEMENT

"Multilingual Subtitle Text Detection and Removal using Video Inpainting"

Our goal is to remove embedded subtitles from the video which could be a source of distraction for the viewers who desire to watch the video without subtitles or for those who do not understand the language of the subtitles.

This project can be used for other applications such as video re-editing and overlaying other subtitles as well.

BACKGROUND

[1] has explained a novel method to extract text using statistical methods. Subtitles being the high contrast objects, can be detected using these methods (preprocessing). [2] built a novel model to detect scene text in natural images. This methodology helped in removing the background noise that is detected along with the subtitle text. [3] has implemented a state of art video inpainting framework to inpaint the subtitle region with good accuracy, great speed and preserving high resolutions. [4] gave an idea on the end to end pipeline to detect, remove and recognize subtitle text with image inpainting technique.

DATASET AND FEATURES

Dataset

The inhouse dataset consists of 200 videos with *English, Hindi, Malayalam* and *Telugu*. We also created a different dataset for text detection which consisted of image frames from videos with around 16K image frames.

Hardware Requirements

A good GPU and sufficient computational units is required for better performance and lower execution time, which decreases by roughly 4 to 5 times. Good cloud services are preferred for storage purposes.

Software Requirements

Python and its library and packages, Pytorch, OpenCV, NumPy, Pandas, Matplotlib.

Constraints and Assumptions

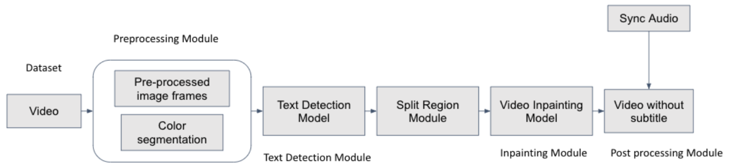
- The entire pipeline has been trained on four languages specifically.
- The videos in the dataset either belong to 720p or 1080p resolution.
- The height of the subtitle region is considered to be a max of 240p.
- We have assumed that the color of the subtitle is white or yellow.

DESIGN APPROACH

An end to end framework is built to detect and remove the subtitle text.

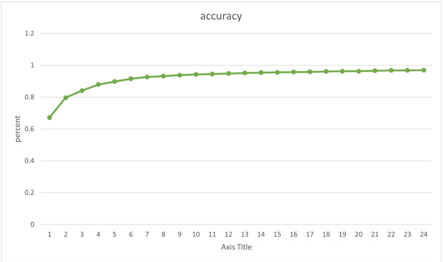
1. The video input is converted to image frames. Each image frame is converted to binary mask by intensity thresholding.
2. A text detection model (CTPN) is built to detect subtitle regions specifically in each image frame.
3. This region is split into sub-regions with dimensions 240 x 432.
4. A video inpainting model (E2FGVI) is employed to fill the subtitle region with appropriate content.

Finally, the inpainted frames are formatted back to a video.

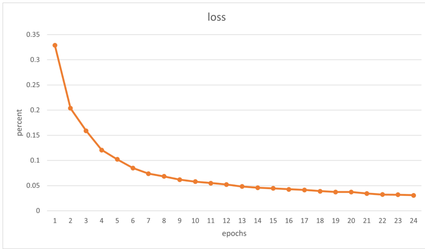


RESULTS AND DISCUSSION

Text Detection Model Train Accuracy plot



Text Detection Train Loss plot



Text Detection

Type	Iou	Precision	Recall	F1-Score
Validation	89.77	91.21	98.30	93.82
Test	90.12	91.88	98.10	94.37

Video Inpainting

Type	Value
PSNR	44.41697732249454
SSIM	0.9175597949261632

SUMMARY OF PROJECT OUTCOME

A subtitle-free video is generated as an output from the framework. A 11 seconds video takes around 2 minutes to be processed by the entire pipeline with a GPU backend (experimented with Tesla T4, Nvidia GEFORCE RTX 3060).

The contributions addressed by our solution are:

- Detecting subtitle text specifically, not scene text.
- Subtitles in all regions (bottom, top, left, right) are detected.
- Preserving the resolution of the video with sub-region inpainting.

CONCLUSIONS AND FUTURE WORK

This project detects and removes subtitles belonging to four languages majorly, English, Hindi, Telugu and Malayalam, as aforementioned. It also provides a decent output on other languages as well (French, Urdu, Russian, Chinese, Bengali).

An improvement can be made in the video inpainting module to consider information from the side-by-side subtitle sub-regions to be more spatially consistent. Also we can consider using KNN module in image inpainting wherein we calculate pixel densities and the K-Nearest Neighboured image frames. This can prevent the usage of a sliding window and the consideration of frames after a certain serial number.

IEEE REFERENCES

[1] Akhtar Jamil, Jawad Rasheed, Bulent Bayram. *"Local statistical features for multilingual artificial text detection from video images"*. 2nd International Conference on Advanced Technologies, Computer Engineering and Science, Alanya, Turkey,2019

[2] Zhi Tian, Weilin Huang, Tong He, Pan He, Yu Qiao. *"Detecting Text in Natural Image with Connectionist Text Proposal Network"*. European Conference on Computer Vision,2016.

[3] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, Ming-Ming Cheng. *"Towards An End-to-End Framework for Flow-Guided Video Inpainting"*. CVPR 2022,April,2022.

[4] Haoran Xu , Yanbai He , Xinya Li , Xiaoying Hu , Chuanyan Hao, Bo Jiang , "*Joint Subtitle Extraction and Frame In-painting for Videos with Burned-In Subtitles*", Special Issue Recent Advances in Video Compression and Coding, 2021.

Name: Ramya C **SRN: PES1UG19CS379** 	**Name: Ratna Bojja** **SRN: PES1UG19CS381** 	**Name: Rishab S** **SRN: PES1UG19CS386** 	**Name: Sahana B Manjunath** **SRN: PES1UG19CS411** 	**Name: Prof. V R Badri Prasad**