Analyzing the Merits and Drawbacks of Multilingual Text Detection Datasets for Videos

Vidya[1], Dr. Manjula G R[2]

[1] Research Scholar,Department of Computer Science& Engg, JNNCE,Shimoga,Karnataka,India
[2] Professor,Department of Computer Science& Engg, JNNCE,Shimoga,Katnataka,India

---***---

Abstract –In recent years, video contents have observed a massive growth with online lecture, news channels, and media contents and with short clips. These contents have objects, texts in them. The majority of the time, multilingual datasets are not available as per the need and have to utilize YouTube videos with multiple languages as a source . This study proposes a systematic structure for the generation of multilingual datasets and explains the purpose for their creation. Additionally, the study demonstrated the difficulties in creating the data set, like segregating data that is multilingual, etc. Lastly, discusses the effects of creating multilingual datasets because of various difficulties. As a result, the research's contribution is the multilingual dataset that was created using the suggested systematic structure, which also practically shows its limitations and benefits.

***Key Words*:**, Data Set, Multilingual, Text detection, challenge, consequences

1.INTRODUCTION

In today's age of information , videos have become a dominant force. They are an effective medium for news dissemination, education, and communication in addition to being a great source of enjoyment. This dominance is the result of multiple variables, like as :-

Engaging Format: Videos capture attention more effectively than text alone, leveraging sound, visuals, and motion to convey information in a compelling way.

Accessibility and Convenience: Videos are readily available on various platforms, accessible from anywhere with an internet connection. They are also easily shared and consumed on mobile devices, further increasing their reach.

Global Appeal: Videos often transcend language barriers, relying on visuals and storytelling to convey meaning.

However, the true power of video lies not just in the moving images but also in the embedded textual information they often contain. This text can take various forms here few of them noted, as Subtitles and Closed Captions, On-Screen Text Overlays, Environmental Text.

For the creation of multilingual video text datasets, where text is treated as an object of interest, represents a pivotal frontier in multimedia research and development. Unlike traditional video datasets, which primarily focus on visual content, these datasets place a special emphasis on capturing and analyzing textual elements within videos across multiple languages. This approach opens up new avenues for understanding and interpreting video content, particularly in scenarios where textual information plays a significant role in conveying meaning or context

Several essential steps must be taken to produce multilingual video text datasets that include text as an object. These are

1. **Data Acquisition**: Gathering a diverse collection of videos that contain prominent textual elements, such as on-screen captions, subtitles, signage, or graphical text overlays. These videos should encompass a wide range of genres, languages, and cultural contexts to ensure the dataset's breadth and representativeness
2. **Text Extraction**: Extracting textual information from the videos using optical character recognition (OCR) or similar techniques. This step involves accurately detecting and parsing text instances within the video frames, accounting for variations in font styles, sizes, colors, and backgrounds
3. **Language Identification**: Identifying the language(s) of the extracted text to facilitate multilingual analysis and annotation. This process may involve language detection algorithms or manual verification by linguists proficient in the target languages
4. **Annotation and Localization**: Annotating the extracted text with relevant metadata, such as its position, size, font type, and language. Additionally, localizing the text by aligning it with the corresponding regions of interest within the video frames to establish spatial and temporal relationships
5. **Quality Assurance**: Conducting thorough quality checks to ensure the accuracy and integrity of the extracted text annotations. This may involve human verification, automated validation algorithms, or a combination of both to mitigate errors and inconsistencies
6. **Dataset Enrichment**: Enriching the dataset with additional contextual information, such as video titles, descriptions, timestamps, and associated metadata. This contextual data enhances the dataset's utility for various applications, including content analysis, retrieval, and understanding

With multilingual video text datasets where text is treated as an object, researchers and practitioners can advance the frontiers of multimedia analysis, enabling more nuanced and comprehensive exploration of textual information within video content. These datasets serve as invaluable resources for developing algorithms and models that can automatically detect, extract, translate, and analyze text across diverse linguistic and cultural contexts, fostering innovation in fields such as content understanding, information retrieval, and cross-lingual multimedia processing. The absence of defined processing principles and criteria makes it extremely difficult for researchers to create a broad multilingual dataset with a wide range of scripts and unlabeled texts.

The motivation for creating a multilingual dataset and proposing a framework stems from the challenges faced, including issues with labeling, annotation, and varied scripts. This research outlines the consequences of these challenges on multilingual dataset creation.

2. CONTENT.

The paper is organized as follows: Section II offers a summary of relevant papers. The challenges and aims addressed are outlined in Section III. A suggested strategy for creating multilingual datasets is provided in Section IV. There is a case study in Section V. Section VI discusses the difficulties in creating a multilingual dataset, and Section VII explores the implications. Section VIII, at last, brings the paper to a close.

3. RELATED WORK

While there might not be research papers directly focused on the specific process of creating multilingual datasets for text detection and extraction in videos, here are relevant papers that explore aspects of this concept:

The research paper [1] introduces a multilingual text dataset (MVTec_Multi-Language) containing text in various languages for text recognition tasks. This dataset structure (text in images) aligns with the concept of extracting text from video frames.and the dataset are for LLM. : "ICDAR2015 Robust Reading Competition" (2015) by Shafait et al.:[2] https://rrc.cvc.uab.es/ is the most common way of formatting the dataset for text detection, extraction and recognition. But the video dataset in this are from single language. The website reference [3] gives a list of multilingual dataset, again they are related to natural language processing where the text are related to one another, they have semantic meanings. Massive data set [4] contains 1M data conversations from 51 language which can assist robot assistance and used to train website assistance mode[5], .Bi-lingual text detection dataset is a dataset containing only two language English and Manipuri in it and these are image datasets[6]. Thomee et al. [7] looked at the process of creating a multimedia data set, concentrating on Flickr photos and the challenges that come with them. They explained why creating a multimedia data set for scientific research was necessary. Their main areas of interest are metadata inadequacy and multimedia content annotation. Consequently, an annotated data collection is produced.The difficulties with multilingual datasets and multilingual data on the web are attempted to be resolved by Gracia et al. [8]. They claim that it could possibly be difficult to convert data stated in many languages into analyzable form, which leads to a large accumulation of useless material that is actually helpful for creating datasets. The dataset for linguistic plausibility was developed by Lambert et al. [9]. They encounter difficulties getting interannotator agreement since the linguistic plausibility are quite subjective. Whiting et al.[10]a dataset for the IEEE Visual Analytics Science and Technology contest was created. The development of visual analytics tools and the adoption of appropriate design presented hurdles when working with this data set.

Given the unmet requirements of available datasets, the pursuit of an in-house multilingual dataset is motivated by the need to fulfill proposed standards and unlock new possibilities in research and application

The University of Oxford's Visual Geometry Group (VGG) created the computer vision tool known as the VGG annotation tool [11]. With the goal to generate labeled datasets for training and assessing machine learning models, it is used to annotate images with different labels, such as object bounding boxes or semantic segmentation masks. This is especially useful in the areas of image recognition and object identification.

4. RESEARCH PROBLEM AND OBJECTIVES

The results of the study make it clear that there isn't a suitable framework for creating multilingual datasets, which highlights the need for creating one. Thus, the following are the goals:
1. To suggest a structure for the production of multilingual datasets
2. To assess the difficulties and effects of creating multilingual datasets

5. METHODOLOGIES

A systematic structure for the generation of multilingual video datasets is suggested to fulfill the goals of this study. The entire procedure is broken down into certain unique, detailed phases. Each step is indispensable; there are no shortcuts.

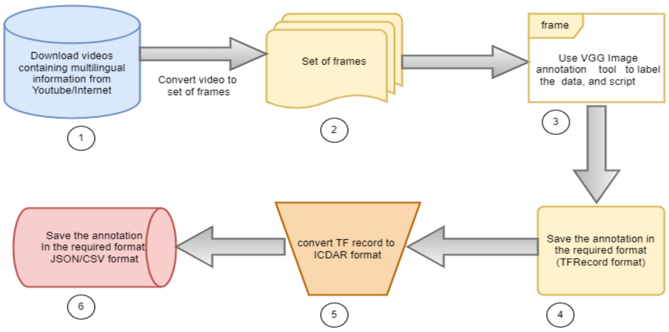

Download videos containing multilingual information from Youtube/Internet
1
Convert video to set of frames
Set of frames
2
frame
Use VGG Image annotation tool to label the data, and script
3
Save the annotation in the required format (TFRecord format)
4
convert TF record to ICDAR format
5
Save the annotation In the required format JSON/CSV format
6

Figure 1: Steps for creating multilingual video dataset in ICDAR format

Following are the steps followed for creating multilingual video frame dataset in ICDAR format.

Step1: Select the videos containing multilingual text from the collection of downloaded videos.

Step2: Using proper code/tool converts them into set of frames. In proposed work python code is written to convert video into set of frames.
Step3: Use VGG tool, select frame/frames to draw bounding box and script labeling [As in Figure 2].
Step4: VGG via tool allows saving annotated images in CSV, JSON, XML format. In the proposed work, downloaded it as CSV file [as shown in figure 3].
Step5: With python code look into the image description field and get the coordinates and convert in to ICDAR format (x1,y1,x2,y2,x3,y3,x4,y4,scipt) [As given in figure 4].
Step6: Depending on the further requirement, save the value as CSV/JSON/TXT file. In the proposed model, they are saved as txt files.

6. RESULTS AND DISCUSSIONS

In order to get the ground truth values required to train the models and to create In-house dataset

VGG Via [11] a web based annotation tool, used to annotate objects, texts, where text are considered as objects. Figure 2, shows how the annotation tool used to annotate bounding boxes for text regions. Once all the text regions are drawn, the result downloaded in comma separated values format, as shown in Figure3.Challenge here is that downloaded result have bounding box value in x,y,height,width format and the required coordinate values for the proposed work is x1,y1,x2,y2,x3,y3,x4,y4. In other words 4 coordinate values is the output we got from annotation tool, and the required format is 8 coordinate values.

With the help of coding with python FOUR coordinate values are converted to EIGHT coordinates. And these set of values further used to train our deep learning model created for multilingual text detection.

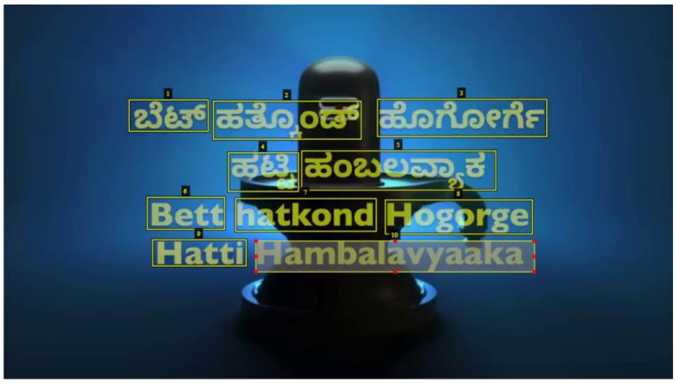
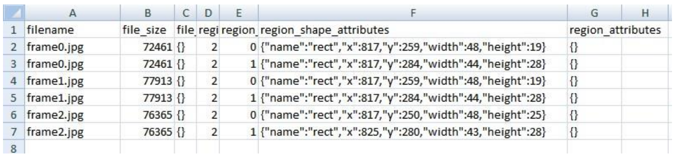
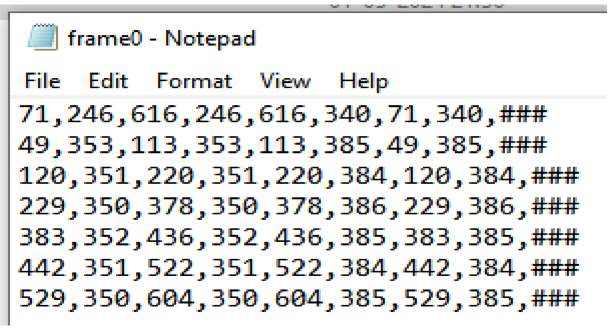


Figure 2: Example frame uploaded and region selected image



	A	B	C	D	E	F	G	H
1	filename	file_size	file_	regi	region_	region_shape_attributes	region_attributes	
2	frame0.jpg	72461	{}	2	0	{"name":"rect","x":817,"y":259,"width":48,"height":19}	{}	
3	frame0.jpg	72461	{}	2	1	{"name":"rect","x":817,"y":284,"width":44,"height":28}	{}	
4	frame1.jpg	77913	{}	2	0	{"name":"rect","x":817,"y":259,"width":48,"height":19}	{}	
5	frame1.jpg	77913	{}	2	1	{"name":"rect","x":817,"y":284,"width":44,"height":28}	{}	
6	frame2.jpg	76365	{}	2	0	{"name":"rect","x":817,"y":250,"width":48,"height":25}	{}	
7	frame2.jpg	76365	{}	2	1	{"name":"rect","x":825,"y":280,"width":43,"height":28}	{}	
8								

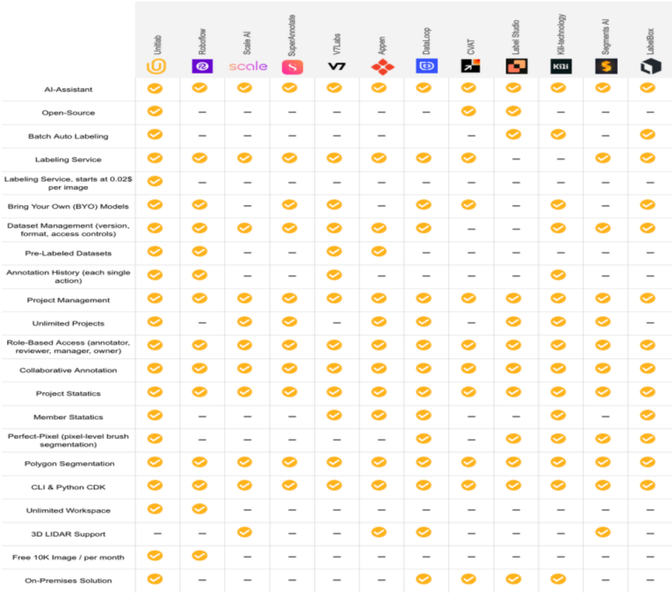
Figure 3: VGG via annotation CSV downloaded file



```
71,246,616,246,616,340,71,340,###
49,353,113,353,113,385,49,385,###
120,351,220,351,220,384,120,384,###
229,350,378,350,378,386,229,386,###
383,352,436,352,436,385,383,385,###
442,351,522,351,522,384,442,384,###
529,350,604,350,604,385,529,385,###
```

Figure 4: Bounding box Text regions in ICDAR format

Following image shows the comparison between different annotation tools suitable for different purposes. Of the tools mentioned in the figure only three are open source and most of them can not able to annotate videos.



	Unitlab	Roboflow	Scale AI	SuperAnnotate	V7Labs	Appen	DataLoop	CVAT	Label Studio	Kili-technology	Segments AI	LabelBox
AI-Assistant	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Open-Source	✓	–	–	–	–	–	–	✓	✓	–	–	–
Batch Auto Labeling	✓	–	–	–	–	–		–	✓	✓	–	✓
Labeling Service	✓	✓	✓	✓	✓	✓	✓	✓	–	–	✓	✓
Labeling Service, starts at 0.02$ per image	✓	–	–	–	–	–	–	–	–	–	–	–
Bring Your Own (BYO) Models	✓	✓	–	✓	✓	–	✓	✓	–	✓	–	✓
Dataset Management (version, format, access controls)	✓	✓	✓	✓	✓	✓	✓	–	–	✓	✓	✓
Pre-Labeled Datasets	✓	✓	–	–	✓	✓	–	–	–	–	–	–
Annotation History (each single action)	✓	✓	–	–	✓	–	–	–	–	✓	–	–
Project Management	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Unlimited Projects	✓	–	✓	✓	–	✓	✓	–	✓	✓	✓	–
Role-Based Access (annotator, reviewer, manager, owner)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Collaborative Annotation	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Project Statatics	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Member Statatics	✓	–	–	–	✓	✓	✓	–	–	✓	–	✓
Perfect-Pixel (pixel-level brush segmentation)	✓	–	–	–	–	–	✓	–	✓	✓	✓	✓
Polygon Segmentation	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CLI & Python CDK	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Unlimited Workspace	✓	✓	–	–	–	–	–	–	–	–	–	–
3D LIDAR Support	–	–	✓	–	–	✓	✓	–	–	–	✓	–
Free 10K Image / per month	✓	✓	–	–	–	–	–	–	–	–	–	–
On-Premises Solution	✓	–	–	–	–	–	✓	✓	✓	✓	–	–

Figure 5: Comparison of different annotation tools

7. CONCLUSIONS

While working on deep learning or machine learning research work , it is difficult to find the datasets readily available when the tasks are challenging , as in proposed work where text is treated like as object. Tools related to text annotation are belongs to natural language processing, where text have relationships with other text in a sentence. Coming with single annotation tool for handling Multilingual text is video is still under research and have to go long way. Proposed work is one such effort to create multilingual dataset, where 20 video clips are considered with a total of 112 frames and prepared them for training the deep learning model of multilingual text detection.

REFERENCES

[1] MVTec_Multi-Language Text Dataset for Text Recognition in Images" (2021) by Baek et al.: https://arxiv.org/pdf/2403.16592

[2] : "ICDAR2015 Robust Reading Competition" (2015) by Shafait et al.: https://rrc.cvc.uab.es/

[3] https://metatext.io/datasets-list/multi-lingual-language

[4] MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages

[5] Veronica, ‘Bi-Lingual Text Detection dataset’. 28-Aug-2023

[6] NISP- A Multi-lingual Multi-accent Dataset for Speaker Profiling

[7]B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, et al., "The new data and new challenges in multimedia research," arXiv preprint arXiv:1503.01817, vol. 1, 2015.

[8] J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Gómez-Pérez, P. Buitelaar, and J. McCrae, "Challenges for the multilingual web of data," Journal of Web Semantics, vol. 11, pp. 63-71, 2012.

[9] B. Lambert, R. Singh, and B. Raj, "Creating a linguistic plausibility dataset with non-expert annotators," in Eleventh Annual Conference of the International Speech Communication Association, 2010.

[10]M. A. Whiting, C. North, A. Endert, J. Scholtz, J. Haack, C. Varley, et al., "VAST contest dataset use in education," in 2009 IEEE Symposium on Visual Analytics Science and Technology, 2009, pp. 115-122.

[11].Abhishek Dutta and Andrew Zisserman. 2019. The VIA Annotation Software for Images, Audio and Video. In Proceedings of the 27th ACM International Conference on Multimedia (MM ’19), October 21–25, 2019, Nice, France. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3343031.3350535