Accurate Name Extraction from News Video Graphics

Andrea Filiberto Lucas
andrea.f.lucas.22@um.edu.mt
University of Malta
Department of AI

Supervisor: Dr. Dylan Seychell
dylan.seychell@um.edu.mt
University of Malta
Department of AI

***Abstract*—As media increasingly relies on video content, automatically extracting crucial information has become essential. Graphical elements like captions and lower thirds often contain key identifiers such as names, but manual extraction is time-consuming and error-prone due to diverse design styles. This report presents the design and implementation of the Accurate Name Extraction Pipeline (ANEP), which leverages YOLO for region detection, Optical Character Recognition (OCR) for text extraction, and Named Entity Recognition (NER) for validation in extracting names from news video graphics.**

ANEP is trained and evaluated using the News Graphics Dataset (NGD), a custom dataset featuring diverse graphical designs from various news sources. The system is compared with two generative AI (GenAI) models, applied to extract names from the same input video, assessing performance through precision, recall, F1-score, and processing time to provide a comprehensive evaluation of traditional versus generative name extraction methods for media analysis and information retrieval.

***Index Terms*—News Video Graphical Analysis, CV, OCR, NER.**

I. Introduction

In the current era of information, the large volume of available data can overwhelm users [1], making it difficult to process and use information effectively, particularly in the media industry, where staying informed is crucial [1]–[3]. The diversity and complexity of the content contribute further to information overload [1], [2]. With news videos becoming a dominant and popular form of media [3], there is an increasing need for tools that can efficiently extract and highlight key information. Without such tools, valuable information can be missed, limiting the audience's ability to analyse and interpret content [4]–[7].

To address these challenges, systems must be developed to automatically process and analyse key components of news videos. While media analysis has advanced [8]–[12], research on automating name extraction from graphical elements remains limited.

News video graphics, such as lower-thirds, captions, and on-screen text, vary greatly in design across sources, creating challenges for automated systems due to their lack of standardisation [13], [14]. The dynamic nature of these videos, with frequently changing text, further complicates the extraction of names [13]. These factors emphasise the need for adaptable, robust solutions that can efficiently analyse and extract insights from diverse video graphics.

A. Proposed Solution

This research develops an automated system to extract names from graphical elements in news videos. YOLO (You Only Look Once), a Convolutional Neural Network (CNN), detects graphical components in video frames. Optical Character Recognition (OCR) extracts text across diverse font styles and resolutions, while Named Entity Recognition (NER) ensures accurate name identification. This process makes up the Accurate Name Extraction Pipeline (ANEP), which employs traditional computational methods for name extraction.

The ANEP output will be compared with results from two generative AI (GenAI) models applied to the same input video. The comparative analysis will evaluate performance through metrics including precision, recall, F1-score, and processing time, providing a comprehensive assessment of traditional and generative name extraction approaches for media analysis and information retrieval.

B. Aims and Objectives

The aim of this research is to develop a system that accurately extracts names from graphical elements in news videos. This system will help users gain insights into the key personalities featured in the videos, which might otherwise be overlooked. In addition, the system will be compared with two GenAI models to assess the relative performance of traditional versus generative methods in name extraction. Ultimately, the research will provide a structured analysis of the extracted names, enhancing media analysis and accessibility.

This will be achieved through the following objectives:

- **Develop the UI** for the ANEP system to allow video uploads, display extracted names, and access report analysis.
- **Build the News Graphics Dataset (NGD)** for training the YOLO model with diverse graphical elements from various news sources.
- **Develop an OCR system** to achieve at least 90% precision in text extraction across different graphical styles and resolutions.
- **Apply NER** to validate text, ensuring 92% precision in identifying names within context.
- **Integrate GenAI API calls** for name extraction, optimizing for minimal latency.
- **Analyse extracted names** from ANEP and two GenAI models, providing a comparative report with names, timestamps, and other performance insights.

II. Background

This chapter outlines the technologies shaping the project's pipeline, addressing challenges in news graphics variability. It highlights YOLO for graphical element detection and region of interest (ROI) identification, OCR for text extraction, and NER for validation and contextualization, while separately examining GenAI as a comparative method for evaluating traditional techniques.

A. News Video Graphics

News video graphics are pivotal in modern broadcasting, conveying critical information such as names, locations, and event titles, making them indispensable for real-time updates [13]–[17]. Ubiquitous across television, online platforms, and social media [16], [17], these graphics simplify narratives and enhance understanding through visual cues [16]. They are particularly effective in live reporting and breaking news, where swift and clear communication is essential. Additionally, their use of universal symbols and visuals helps overcome language barriers, reaching a global audience [13], [14].

Despite these advantages, news video graphics pose significant challenges for automated systems. The lack of design standardization [18], [19] leads to variations in styles, layouts, and formats across outlets, complicating consistent detection and extraction of relevant information [18]. Furthermore, their dynamic nature, with moving or changing text, adds to the complexity.

B. Identifying Visual Elements in News Videos

The first step in automating name extraction involves detecting graphical elements containing names using a CNN-based model [20]. YOLO is utilized due to its ability to combine localization and classification in a single operation, enabling real-time, accurate detection. It divides images into grids and predicts bounding boxes and class probabilities simultaneously, making it suitable for dynamic content such as news broadcasts [21]–[23]. Advanced YOLO versions improve detection of smaller objects and resolve overlapping items [24], [25].

To develop a robust YOLO-based system, a high-quality annotated dataset is required, capturing the diverse visual styles in news graphics, including variations in fonts, colours, and layouts. The dataset should also account for challenges such as motion blur, overlapping objects, and lighting changes. Data augmentation techniques like rotations, scaling, and brightness adjustments can help train the model for improved accuracy [26].

C. Extracting Text from News Videos Graphics

OCR converts text in images into machine-readable format [27], [28]. The process begins with pre-processing, using techniques such as binarization, contrast adjustment, and edge detection to reduce noise and enhance text clarity [28]. Text is then segmented into characters or words for feature extraction and classification [27].

OCR accuracy can be affected by challenges like font variations, low contrast, motion blur, and visually similar characters [29]. To address these issues, post-processing techniques, including dictionary-based corrections, refine the extracted text and improve contextual coherence [27], [29].

D. Validating Extracted Text with NER

NER, a subset of Natural Language Processing (NLP), validates text extracted by OCR to ensure only names from news video graphical elements are retained [30], [31]. Modern NER systems often use advanced machine learning techniques and pre-trained language models, to handle the complexity and variability of real-world data [31], [32]. These models analyse syntax and semantics to accurately identify names, even in unconventional formats [31].

NER mitigates OCR challenges, mentioned in II-C, by cross-referencing extracted text with linguistic patterns and predefined entity categories [33]. Dictionary-based corrections and contextual analysis further enhance accuracy in identifying names [34].

E. Using GenAI for Name Extraction

GenAI provides a novel approach to extracting names from video content. Models like Generative Adversarial Networks (GANs) and Variational Auto-encoders (VAEs) are well-suited for noisy or distorted data, enabling name detection and extraction even when text is partially obscured or degraded [35]–[37].

Despite their advantages, GenAI models face challenges, particularly regarding training data quality and diversity. Biases or inaccuracies in datasets can affect output reliability, while ensuring model explainability is critical for verifying accuracy in sensitive applications like name extraction from news videos [36], [37].

III. Literature Review

This chapter reviews the literature on extracting information from videos, with a focus on object detection and text extraction. Key advances in CV, OCR, and NER are discussed, emphasizing their significance as the foundation for developing the proposed ANEP system. The review also explores recent work on generative AI models for text extraction, aiming to evaluate and compare their performance against conventional methods

A. CV Techniques for Video Frame Processing

Preethi et al. [39] emphasize the importance of object detection algorithms such as YOLO and Faster R-CNN, each offering distinct advantages. YOLO processes images in a single pass, balancing speed and accuracy, making it ideal for real-time applications. In contrast, Faster R-CNN employs a two-stage process where the Region Proposal Network (RPN) generates object locations, refined through classification and regression, excelling in precision for intricate or occluded objects [39]. H. and Venkatapur [40] underscore the trade-off between speed and accuracy when selecting between these methods.

YOLO divides an image into a grid, predicting bounding boxes and class probabilities for objects in each cell. Anchor boxes anticipate object shapes and sizes, while non-maximum suppression (NMS) filters redundant boxes, leaving the most probable ones. Hammoudeh et al. [41] note this efficient approach and recommend exploring YOLOv8, which incorporates attention mechanisms and adaptive loss functions for enhanced detection capabilities [41].

Faster R-CNN combines RPN and Fast R-CNN detectors, using shared convolutional layers for efficiency. ROI pooling refines classifications and bounding boxes by analysing region proposals from the RPN with objectivity scores. Despite its complexity, it offers superior accuracy, particularly in low-resolution or cluttered video frames [39], with feature pyramid networks (FPN) further improving performance for fine-grained detection.

B. OCR in Video Analysis

OCR plays a critical role in extracting textual information from videos, requiring advanced techniques to tackle challenges presented by dynamic video frames. Tesseract, discussed by Ashraf et al. [42] and Smith [43], exemplifies significant advancements in OCR technology. Initially developed as a research project at HP Labs and later enhanced by Google, Tesseract uses a pipeline based on connected component analysis to detect and recognize text, including inverted text and diverse fonts. However, applying OCR to video frames introduces issues such as motion blur, varying resolutions, and noise. To address these challenges, pre-processing methods like gray-scale conversion, binary thresholding, and noise reduction are essential. These techniques optimize the quality of video frames before analysis, thereby improving OCR accuracy [42].

Smith [43] discusses Tesseract's segmentation strategy, which combines initial line finding with adaptive classifiers that distinguish character cases through baseline/x-height normalization and reduce distortions via moment normalization. While these strategies perform well in static settings, they struggle in dynamic video contexts, where text outlines can become inconsistent due to compression artifacts or rapid movement. To overcome this, methods like Hidden Markov Models (HMMs) or Multi-dimensional Long Short-Term Memory (MDLSTM) networks have been proposed to enhance sequential character recognition [42]. Smith [43] also notes that Tesseract's polygonal approximations may underperform in dynamic settings, suggesting the need for more robust segmentation techniques tailored to video analysis.

Combining OCR with video-specific pre-processing and enhancements offers promise for improving text extraction. Ashraf et al. [42] propose video-specific pipelines, which include normalization and noise handling to prepare frames for OCR processing. Additionally, Smith [43] emphasizes the importance of adaptive learning models that dynamically refine recognition based on contextual changes in video. These ongoing improvements, such as Tesseract's chopping mechanism and character n-gram models, aim to bridge the gap between OCR's current capabilities and the demanding requirements of video analysis.

C. Extracting Names Using NER

NER plays a crucial role in extracting structured information, such as names and entities, where it enhances information retrieval. [44] emphasizes the importance of NER in these contexts, discussing rule-based approaches that rely on predefined linguistic patterns and domain knowledge. While effective in consistent domains, rule-based methods lack adaptability in more dynamic contexts. In contrast, learning-based methods such as Conditional Random Fields (CRFs) and Long Short-Term Memory (LSTM) networks excel at contextual understanding and adaptation [44]. Naseer et al. [45] propose hybrid models that combine rule-based precision with the adaptability of machine learning, highlighting CRFs' ability to model sequential dependencies. This is especially crucial for tasks like entity boundary detection in noisy data [45].

Emerging models like spaCy and BERT show significant promise for video-specific NER. Chavan and Patil [44] highlight spaCy's efficiency and its ease of fine-tuning for domain-specific applications, making it suitable for video text extraction. Pakhale [46] argues that BERT's transformer architecture is particularly well-suited for video NER due to its ability to capture contextual nuances and resolve ambiguities in noisy or short text segments. Fine-tuning these models on video-specific datasets could address common OCR issues such as errors and limited context, suggesting a promising future for NER in video analysis. However, substantial adaptation and training efforts are still required to fully realize the potential of these models in this domain.

D. Generative AI for Name Extraction

Traditional OCR methods, as discussed in sectionr III-B, struggle with noise, distortions, and text style variations. In contrast, transformer-based models like TrOCR offer key advantages. By leveraging pre-trained models such as DeiT, BEiT, and RoBERTa, TrOCR benefits from large datasets for improved feature extraction and language modelling [47]. Its end-to-end encoder-decoder architecture processes image patches as sequences, eliminating the need for CNN backbones [47]. TrOCR also generates text at the word-piece level, reducing reliance on external language models.

Transformers excel at contextual understanding, helping distinguish names from other text elements by analysing linguistic and visual cues. They focus on visual markers like font, size, and position, aiding name identification near faces or in graphic elements like lower thirds. By linking text with surrounding visuals, transformers enhance name extraction and provide deeper semantic insights. Their attention mechanisms make transformers resilient to noise, distortion, and occlusions, making them ideal for dynamic video environments where text may shift, move, or partially obscure. As shown in [48], transformers excel at multimodal analysis, tracking names in video graphics and maintaining high accuracy even under challenging conditions.

IV. Design

This chapter provides a high-level overview of the system's design, detailing the workflow and interactions among the core components. The rationale behind each component is discussed, linking the design choices to the literature reviewed in Chapter III, with the aim of addressing the challenges associated with this project.

A. Overview

As shown in figure 1, the system is designed to accurately extract names from news videos. The approach involves using a custom pipeline for name extraction, which compares the results against two GenAI models. The system consists of several interconnected components working together to achieve this goal.

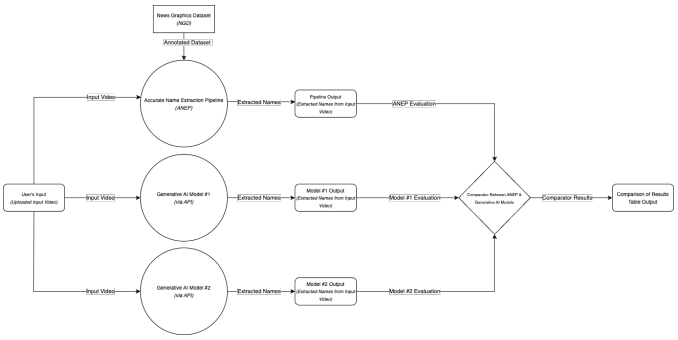

Fig. 1: DFD Level 1

The process begins with the user providing an input video, which is then processed by the ANEP. The pipeline uses a pre-trained YOLO model, trained on the custom NGD, to detect graphical elements in video frames. OCR extracts the text from these regions, and NER validates and identifies names. The final output is a set of extracted and validated names, presented to the user. In parallel, the input video is also sent to two GenAI models via APIs, allowing them to extract names from the video. The system compares its results with those of the GenAI models, focusing on performance metrics such as precision, recall, F1-score, and processing time.

B. Components

The system begins with the User Interaction component, where users upload .mp4 news videos for processing. The input video is processed by the Accurate Name Extraction Pipeline (ANEP) and also sent to GenAI models via APIs for name extraction.

The ANEP includes two sub-components: Graphics Detection, which uses a YOLO model trained on the NGD to locate graphical elements in video frames, and Text Extraction, where OCR extracts text and NER identifies and validates names.

Finally, the Results Visualization component presents the extracted names, timestamps, locations, screenshots, and a metric comparison between the traditional pipeline and GenAI models. Figure 2 shows the GenAI Interactions, whilst figure 3 illustrates the ANEP Interactions.

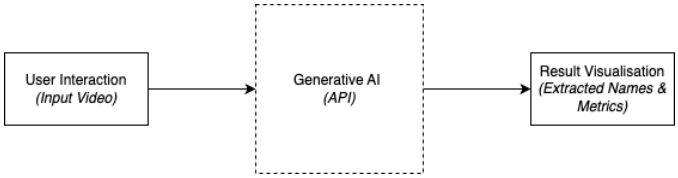

Fig. 2: GenAI Component Interactions

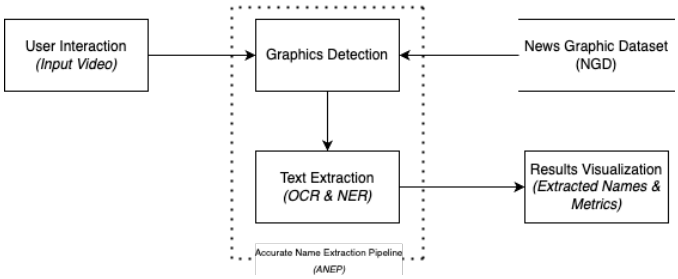

Fig. 3: ANEP Component Interactions

V. Implementation

This chapter outlines the system's implementation, building on Chapter IV, with a high-level overview of its structure and functions, followed by the implementation of individual components, including prototype algorithms, tools, and libraries. It mainly focuses on the ANEP (traditional pipeline) methodology, as the GenAI models are simply integrated via API calls, for name extraction.

A. System Overview

The system's implementation enables the analysis of uploaded .mp4 news videos, with the primary goal of accurately extracting and presenting names found within the video content. The workflow starts with video upload and ends with a detailed output, including extracted names, timestamps, positional data, cropped screenshots, and a comparative analysis of results from the ANEP and GenAI models.

The architecture is divided into three main components: **user interaction**, the **ANEP**, and **results visualization**. ANEP comprises two sub-components: **graphics detection**, using a YOLO model trained on the NGD to locate graphical elements, and **text extraction**, employing OCR and NER to extract and validate names. Input videos are also processed by GenAI models via API integration.

Python is used for the ANEP due to its robust libraries for object detection, OCR, and NER. The system is built on a Python Flask back-end for data handling.

This modular and scalable architecture supports future enhancements and highlights the system's practical utility, aligning with the project's objectives outlined in section I-B.

B. User Interactions

The User Interaction component provides a streamlined UI for users to upload an .mp4 news video and initiate the system. The interface includes a central drag-and-drop area,

supported by a clickable option for file selection. Once a video is successfully uploaded, users receive visual feedback indicating the file name and upload status, followed by the activation of a ”Run Analysis” button. Upon initiation, the interface transitions to a progress screen displaying a loading indicator, estimated processing time, and a cancel option for user control. Post-analysis, the interface will present a visualization of model comparison metrics and extracted name analytics, offering an intuitive workflow for user interaction.

C. News Graphic Dataset (NGD)

The NGD is a curated collection of annotated frames extracted from news videos sourced online from local and foreign news media outlets. These frames highlight graphical elements, annotated with bounding boxes using Roboflow, to prepare them for training the YOLO model. Video segments are sourced from platforms like the Internet Archive's TV archive and YouTube, using tools such as 'yt-dlp' for high-quality downloads. Figure 4 shows samples of annotations from various news sources, with the graphic elements visibly highlighted.

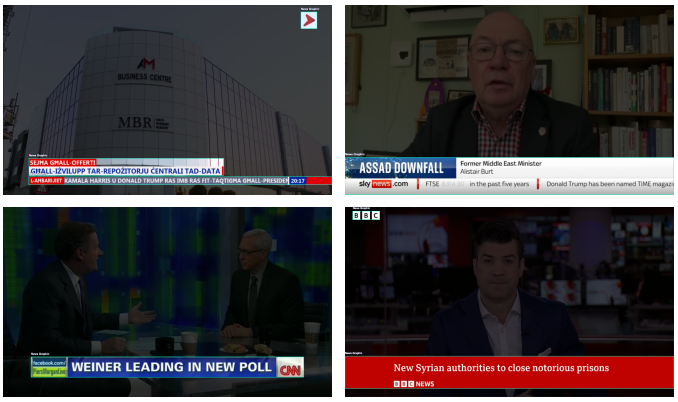

Fig. 4: Sample annotations from different news sources

D. Graphics Detection

The Graphical Detection module, the first sub-component of the ANEP, identifies regions containing graphical elements in video frames using a custom-trained YOLO model. This model is trained on annotated video samples from the NGD, enabling it to learn features and patterns of graphical elements in news videos. Through supervised learning, the YOLO model is exposed to labelled frames to detect and localize textual elements. The trained model efficiently identifies areas for further name extraction processing within the ANEP system.

E. Text Extraction - OCR

The Text Extraction component, the second and most crucial sub-component of the ANEP, identifies and extracts text from video frames using OpenCV and Tesseract. OpenCV processes video frames in real-time, applying pre-processing techniques like gray-scaling and thresholding to improve text visibility by reducing noise and enhancing contrast. The Tesseract engine, accessed via `pytesseract.image_to_data()`, extracts detailed text information, including content and position, even in noisy or complex backgrounds. This combination ensures accurate and reliable text extraction for the ANEP system.

F. Text Extraction - NER

After text extraction, the system uses `spaCy`'s pre-trained `en_core_web_md` model for NER to identify "PERSON" entities, representing potential names. This helps distinguish relevant names from other text. Moreover, a regular expression (regex) filters out URLs, ensuring only valid names remain. To optimize performance, the system employs Python's `ThreadPoolExecutor` for multi-threading, processing multiple frames concurrently to handle large video files efficiently.

```
Processing frames:  97.5%                                  3121/3202 frames
No more frames to read or video is over.

Detected Names with Timestamps and Positions from Video
+-----------------+-----------------+-----------------------------------+
| Timestamp (s)   | Detected Name   | Position (x, y, width, height)    |
+-----------------+-----------------+-----------------------------------+
|          7.24   | Keir Starmer    | (x=90, y=997, width=1830, height=45)  |
|          7.24   | Donald Trump    | (x=90, y=997, width=1830, height=45)  |
|          7.24   | Keir Starmer    | (x=90, y=997, width=1827, height=45)  |
|          7.24   | Donald Trump    | (x=90, y=993, width=63, height=53)    |
|          7.24   | Keir Starmer    | (x=90, y=993, width=1811, height=53)  |
|          7.24   | Keir Starmer    | (x=90, y=997, width=1830, height=45)  |
|          7.96   | Keir Starmer    | (x=356, y=0, width=758, height=1080)  |
|          8.40   | Keir Starmer    | (x=90, y=997, width=1829, height=45)  |
|          8.44   | Donald Trump    | (x=90, y=997, width=1829, height=45)  |
```

Fig. 5: ANEP Sample: Extracted Names from Sample Video

G. Results Visualization

After text extraction and validation, the identified names from the video will be compiled into CSV files, including the names, timestamps, frame locations, and relevant screenshots. This structured file allows the user to easily review, analyse, and process the extracted named entities. The Results Visualization component will also generate statistics, such as the total number of names, unique names, and name distribution across video segments, providing an overview of the extraction process. Additionally, the evaluation and comparison of the custom pipeline with two GenAI models will be presented in a table, as shown in Table I.

TABLE I: Sample: Models Evaluation Metrics

Models	Precision (%)	Recall (%)	F1-Score (%)	Time (s)
ANEP				
GenAI #1				
GenAI #2				

VI. Evaluation

This chapter evaluates the performance of the ANEP system by examining two key components: **graphical element detection** and **name extraction accuracy**. The first sub-section focuses on the system's ability to identify and locate news video graphics within video frames, while the second assesses the accuracy of name extraction and compares it with the two GenAI models, considering both name detection and processing time.

A. Evaluating the Graphical Element Detection

The graphical element detection by the YOLO model will be evaluated using the Intersection over Union (IoU) metric, which measures the overlap between predicted and ground truth bounding boxes. The IoU is calculated by dividing the overlap area by the union area of the bounding boxes. A ground truth dataset, the NGD, will be created, as referenced in section V, containing bounding box annotations for graphical elements in the news videos. The predicted bounding boxes from YOLO will be compared to the ground truth to calculate the IoU and assess detection accuracy.

B. Evaluating the extracted Names

The evaluation of extracted names will use Precision, Recall, and F1-score. **Precision** measures the proportion of true positives among all identified names, while **Recall** evaluates the proportion of true positives among all ground truth names. The **F1-score** combines Precision and Recall, providing a balanced measure of accuracy and completeness.

C. Evaluation of the Three Different Models

In this section, the performance of the traditional ANEP model will be compared with two generative AI models. The evaluation will focus on the extracted names' Precision, Recall, and F1-score for each model. Additionally, the time taken for each model to complete the name extraction process will be measured and compared to assess the efficiency of the models. The results will be outputted in a table, as shown in Table I.

VII. Conclusion

This report outlines the preliminary research, design, and implementation of the ANEP system, aimed at accurately extracting and validating names from news video graphics in a fully automated manner. The research addresses challenges in video content processing, including graphical element detection with YOLO and text extraction with OCR and NER. The proposed architecture is scalable and adaptable, with methods for evaluation and performance assessment. Future work includes further optimization of the pipeline, choosing the most suitable GenAI models, comparing them with the traditional model, and improving the system's UI.

References

[1] P. Aussu, "Information Overload: Coping Mechanisms and Tools Impact," in Lecture notes in business information processing, 2023, pp. 661–669. doi: 10.1007/978-3-031-33080-3_49.

[2] P. G. Roetzel, "Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development," BuR - Business Research, vol. 12, no. 2, pp. 479–522, Jul. 2018, doi: 10.1007/s40685-018-0069-z.

[3] A. A. Naz and R. A. Akbar, "Use of media for effective instruction its importance: some consideration," Journal of elementary education, vol. 18, no. 1-2, pp. 35–40, 2008.

[4] K. Anoop, M. P. Gangan, and V. L. Lajish, "Mathematical Morphology and Region Clustering Based Text Information Extraction from Malayalam News Videos," in Advances in intelligent systems and computing, 2015, pp. 431–442. doi: 10.1007/978-3-319-28658-7_37.

[5] M. D. A. Asif, et al., "A novel hybrid method for text detection and extraction from news videos," 2014. doi: 10.5829/idosi.mejsr.2014.19.5.21019.

[6] X. Ma, "Research on Short News Video Transmission in the Fusion Media Environment," Probe - Media and Communication Studies, vol. 2, no. 2, p. 25, Feb. 2020, doi: 10.18686/mcs.v2i2.1295.

[7] C. Zachlod, et al., "Analytics of social media data – State of characteristics and application," Journal of Business Research, vol. 144, pp. 1064–1076, Feb. 2022, doi: 10.1016/j.jbusres.2022.02.016.

[8] K. Choro´ s, "Video structure analysis and content-based indexing in the automatic video indexer avi," in Advances in Multimedia and Network Information System Technologies. Springer, 2010, pp. 79–90.

[9] S. Lee and K. Jo, "Strategy for automatic person indexing and retrieval system in news interview video sequences," in 2017 10th International Conference on Human System Interactions (HSI). IEEE, 2017, pp. 212–215.

[10] H. Zhang and Y. Gong, "Automatic parsing of news video," in Proc. IEEE Int. Conf. Multimedia Comput. Syst., Boston, MA, USA, 1994, pp. 45–54.

[11] M. Ghoniem, D. Luo, J. Yang, and W. Ribarsky, NewsLab: Exploratory Broadcast News video analysis. 2007, pp. 123–130. doi: 10.1109/vast.2007.4389005.

[12] M. S. Pattichis, V. Jatla, and A. E. U. Cerna, "A Review of Machine Learning Methods Applied to Video Analysis Systems," arXiv (Cornell University), Jan. 2023, doi: 10.48550/arxiv.2312.05352.

[13] O. Järvi and Department of communication studies, University of Vaasa, Finland, "News Graphics: Some Typological and Textual Aspects," journal-article, 2001.

[14] J. R. Fox, A. Lang, Y. Chung, S. Lee, N. Schwartz, and D. Potter, "Picture This: Effects of Graphics on the Processing of Television News," Journal of Broadcasting & Electronic Media, vol. 48, no. 4, pp. 646–674, Dec. 2004, doi: 10.1207/s15506878jobem4804_7.

[15] J. S. Foote and A. C. Saunders, "Graphic Forms in Network Television News," Journalism Quarterly, vol. 67, no. 3, pp. 501–507, Sep. 1990, doi: 10.1177/107769909006700304.

[16] L. Buijs, Y. de Haan, and G. Smit, "Using Information Visualization in the Media," 2013.

[17] S. Li and P. Jongbin, "The Impact of Social Media on Visual Communication Design," Journal of New Media and Economics., vol. 1, no. 2, pp. 138–145, Mar. 2024, doi: 10.62517/jnme.202410223.

[18] R. Borgo et al., "A Survey on Video-based Graphics and Video Visualization," Eurographics, pp. 1–23, Jan. 2011, doi: 10.2312/eg2011/stars/001-023.

[19] N. H. Li, D. Doermann, and O. Kia, "Automatic text detection and tracking in digital video," IEEE Transactions on Image Processing, vol. 9, no. 1, pp. 147–156, Jan. 2000, doi: 10.1109/83.817607.

[20] I. Aljarrah and D. Mohammad, Video content analysis using convolutional neural networks, vol. 1. 2018, pp. 122–126. doi: 10.1109/iacs.2018.8355453.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection. 2016, pp. 779–788. doi: 10.1109/cvpr.2016.91.

[22] G. Lavanya and S. D. Pande, "Enhancing Real-time Object Detection with YOLO Algorithm," EAI Endorsed Transactions on Internet of Things, vol. 10, Dec. 2023, doi: 10.4108/eetiot.4541.

[23] Y. Zhao and S. Wang, "Research on real-time object detection based on Yolo algorithm," Highlights in Science Engineering and Technology, vol. 7, pp. 323–331, Aug. 2022, doi: 10.54097/hset.v7i.1091.

[24] S. Gothane, "A Practice for Object Detection Using YOLO Algorithm," International Journal of Scientific Research in Computer Science Engineering and Information Technology, pp. 268–272, Apr. 2021, doi: 10.32628/cseit217249.

[25] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," Machine Learning and Knowledge Extraction, vol. 5, no. 4, pp. 1680–1716, Nov. 2023, doi: 10.3390/make5040083.

[26] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," Journal of Big Data, vol. 6, no. 1, Jul. 2019, doi: 10.1186/s40537-019-0197-0.

[27] K. Hamad and M. Kaya, "A Detailed Analysis of Optical Character Recognition Technology," International Journal of Applied Mathematics Electronics and Computers, vol. 4, no. Special Issue-1, p. 244, Dec. 2016, doi: 10.18100/ijamec.270374.

[28] S. Amar, “A Detailed study and recent research on OCR,” www.academia.edu, Jan. 2021, Available: https://www.academia.edu/45351532
[29] J. Wang, “A Study of The OCR Development History and Directions of Development,” Highlights in Science Engineering and Technology, vol. 72, pp. 409–415, Dec. 2023, doi: 10.54097/bm665j77.
[30] B. Mohit, “Named Entity Recognition,” in Theory and applications of natural language processing, 2014, pp. 221–245. doi: 10.1007/978-3-642-45358-8_7.
[31] “A Brief History of Named Entity Recognition.” Available: https://arxiv.org/html/2411.05057v1
[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv (Cornell University), Jan. 2018, doi: 10.48550/arxiv.1810.04805.
[33] E. Arslan, “Natural Language Processing: Named Entity Recognition (NER)” Medium, Nov. 21, 2024. Available: https://medium.com/@erhan_arslan/natural-language-processing-named-entity-recognition-ner-step-4-3c079c878332
[34] K. Nastou, M. Koutrouli, S. Pyysalo, and L. J. Jensen, “Improving dictionary-based named entity recognition with deep learning,” Bioinformatics, vol. 40, no. Supplement_2, pp. ii45–ii52, Sep. 2024, doi: 10.1093/bioinformatics/btae402.
[35] Y. Ye et al., “Generative AI for visualization: State of the art and future directions,” *Elsevier B.V.*, May 2024. doi: 10.1016/j.visinf.2024.04.003.
[36] T. S. Musalamadugu and H. Kannan, “Generative AI for medical imaging analysis and applications,” *Future Medicine AI*, Sep. 2023. doi: 10.2217/fmai-2023-0004.
[37] ”Comparative study on AI and OCR for data extraction,” ThirdEyeData, 2024. [Online]. Available: https://thirdeyedata.ai/comparative-study-on-ai-and-ocr-for-data-extraction/. [Accessed: Dec. 11, 2024].
[38] R. C. Basole and T. Major, “Generative AI for Visualization: Opportunities and Challenges,” *IEEE Computer Graphics and Applications*, vol. 44, no. 2, pp. 55–64, Mar. 2024. doi: 10.1109/mcg.2024.3362168.
[39] G. Preethi, S. Naik, M. Kenchol, S. P. Jakalannanavar, and M. S. Rachana, ”Object Detection Using FasterRCNN, YOLOV7 & YOLOV8,” Indiana Journal of Multidisciplinary Research, vol. 04, no. 03, pp. 136–141, Jun. 2024, doi: 10.5281/zenodo.12674762.
[40] K. P. H and R. B. Venkatapur, ”Deep Learning Technique for Object Detection from Panoramic Video Frames,” International Journal of Computer Theory and Engineering, vol. 14, no. 1, pp. 20–26, Jan. 2022, doi: 10.7763/ijcte.2022.v14.1306.
[41] M. A. A. Hammoudeh, M. Alsaykhan, R. Alsalameh, and N. Althwaibi, ”Computer Vision: A Review of Detecting Objects in Videos – Challenges and Techniques,” International Journal of Online and Biomedical Engineering (iJOE), vol. 18, no. 01, pp. 15–27, Jan. 2022, doi: 10.3991/ijoe.v18i01.27577.
[42] N. Ashraf, S. Y. Arafat, and M. J. Iqbal, ”An Analysis of Optical Character Recognition (OCR) Methods,” International Journal of Computational Linguistics Research, vol. 10, no. 3, pp. 81-91, Sep. 2019. doi: 10.6025/jcl/2019/10/3/81-91.
[43] R. Smith and Google Inc., An Overview of the Tesseract OCR Engine. Available: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/33418.pdf.
[44] T. Chavan and S. Patil, ”Named Entity Recognition (NER) For News Articles,” International Journal of Artificial Intelligence Research and Development (IJAIRD), vol. 2, no. 1, pp. 103–112, Season-01 2024. Available: https://iaeme.com/Home/issue/IJAIRD?Volume=2&Issue=1.
[45] S. Naseer, M. Ghafoor, S. B. K. Alvi, A. Kiran, and S. U. Rehman, ”Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance,” University of the Punjab, 2022.
[46] K. Pakhale, ”Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges,” arXiv preprint, arXiv:2309.14084, Sep. 2023. Available: https://arxiv.org/abs/2309.14084.
[47] M. Li, Y. Liu, and Z. Zhang, ”TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models,” 2021.
[48] H. Akbari, M. Li, and X. Chen, ”VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio, and Text,” 2021.

FYP - Gantt Chart