L-Università ta' Malta
Faculty of Information & Communication Technology
Department of Artificial Intelligence

B.Sc. IT (Hons.)
Artificial Intelligence

Accurate Name Extraction from News Video Graphics

Andrea Filiberto Lucas
Supervisor: Dr. Dylan Seychell

INTRODUCTION

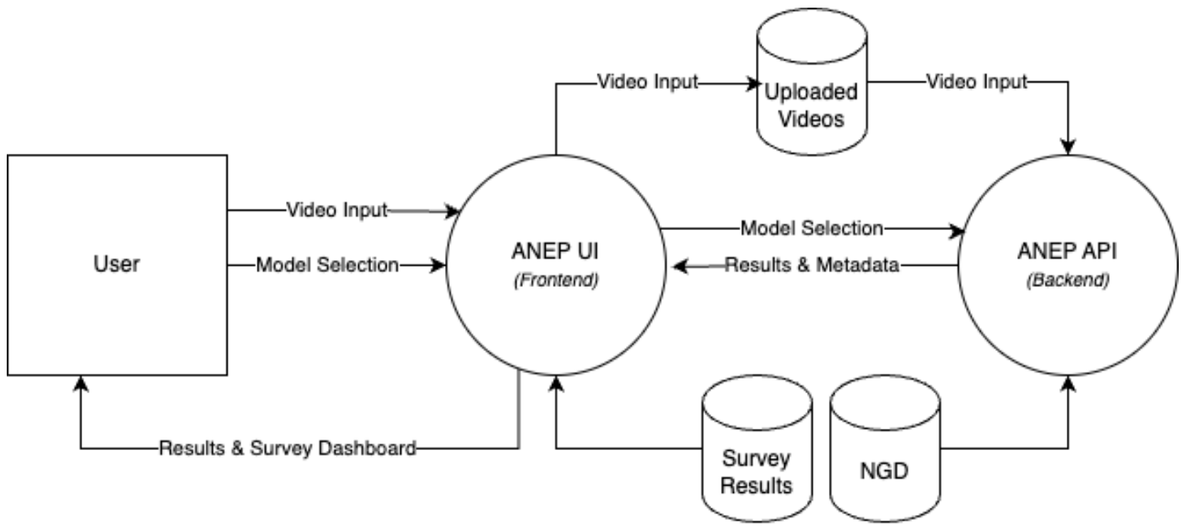
News broadcasts feature a wide variety of graphic overlays, from headlines and lower thirds to scrolling tickers. These elements differ significantly in typography, placement, background designs, and transparency. Each broadcaster employs unique visual styles, making manual extraction of names a laborious task prone to errors. Current automated methods [1] often struggle to handle this visual diversity, leading to frequent misidentification of text or an inability to adapt to the differing design conventions.

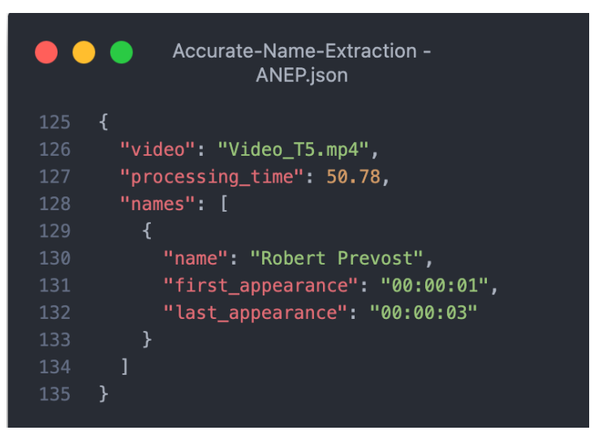
This dissertation introduces the **Accurate Name Extraction Pipeline (ANEP)**, a robust system for extracting names from news graphics. ANEP seamlessly integrates three core technologies: **YOLOv12 for graphic detection**, optical character recognition (**OCR**) for text extraction, and named entity recognition (**NER**) for identifying names. The pipeline's performance is benchmarked against leading Generative AI (**GenAI**) solutions, namely **Google Vision** with **Gemini 1.5 Pro** and **LLaMA 4 Maverick**, offering comprehensive insights into its accuracy and practical applicability.

AIM

This study pursued three primary objectives. Firstly, the **News Graphics Dataset (NGD**) was created, comprising 1,500 frames from local and international news broadcasts as well as social media-based sources. This dataset showcases a diverse range of fonts, layouts, and colour schemes, enabling robust training of YOLOv12 on real-world graphics. Secondly, the **ANEP** was introduced, a pipeline that combines CNN-based Yolov12 detection, OCR for text extraction, and transformer-based NER to precisely locate and validate names. Thirdly, a **comparative framework** was established to benchmark **ANEP against the aforementioned GenAI** methods. Performance metrics, including precision, recall, F1-score, and average runtime, were evaluated to understand the trade-offs between traditional and multimodal approaches.

ARCHITECTURE DESIGN





METHODOLOGY

Developed the NGD by gathering & annotating 1,500 frames → Trained a YOLOv12 model to detect various news graphics → Applied OCR & NER to extract names from detected regions of interest → Compared ANEP with Gemini 1.5 Pro & LLaMA 4 Maverick

RESULTS

Results demonstrated that the locally trained **YOLOv12 model achieved 95.8% object detection accuracy** on the NGD, ensuring reliable localisation of graphical elements across diverse news video frames, even from news sources that were not part of the NGD. Building upon this foundation, the ANEP achieved balanced name extraction performance, with 72.92% precision and 74.44% recall, resulting in **an F1 score of 68.10%.** The system's modular architecture **provided robust and interpretable results**. In comparison, Google Vision integrated with Gemini 1.5 Pro recorded **the highest F1 score of 82.22%,** with superior metrics (93.33% precision, 76.67% recall) and significantly faster processing. However, the LLaMA 4 Maverick pipeline showed weaker performance, achieving **an F1 score of 55.56%.** These findings highlight ANEP's consistency and interpretability for applications requiring audit trails, while also confirming the superior speed and accuracy of GenAI approaches in contexts where transparency requirements are less stringent.

CONCLUSIONS AND FUTURE WORK

This research advances automated name extraction from news video graphics by developing the NGD and ANEP. The outlined approach builds upon earlier television broadcast summarisation systems [2], introducing a generalisable, visual-first methodology for entity extraction while aligning with emerging multimodal information extraction research [3]. Future directions include expanding the NGD to incorporate multilingual and non-Latin based content, optimising ANEP for real-time deployment, and exploring audio-visual integration to enhance contextual understanding.

REFERENCES

[1] J. Hong *et al.*, "Analysis of faces in a decade of US cable TV news," in *Proc. 27th ACM SIGKDD Int. Conference* Knowledge Discovery & Data MiningAssoc. Computing Machinery, 2021

[2] J. Attard and D. Seychell, Comparative analysis of image, video, and audio classifiers for automated news video segmentation, 2025. arXiv: 2503.21848 [Online]. Available: https://arxiv.org/abs/2503.21848

[3] R. Sapkota, S. Raza, M. Shoman, A. Paudel, and M. Karkee, "Multimodal large language models for image, text, and speech data augmentation: A survey", arXiv preprint arXiv:2501.18648, 2025. [Online]. Available: https://arxiv.org/abs/2501.18648