AI as a Tool for Fair Journalism: Case Studies from Malta

Dylan Seychell
Dept. of AI
University of Malta
Msida, Malta
dylan.seychell@ieee.org

Gabriel Hili
Dept. of AI
University of Malta
Msida, Malta
gabriel.hili@um.edu.mt

Jonathan Attard
Dept. of AI
University of Malta
Msida, Malta
jonathan.attard.20@um.edu.mt

Konstaninos Makantatis
Dept. of AI
University of Malta
Msida, Malta
konstantinos.makantasis@um.edu.mt

Abstract—In today's media landscape, the role of Artificial Intelligence (AI) in shaping societal perspectives and journalistic integrity is becoming increasingly apparent. This paper presents two case studies centred on Malta's media market featuring technical novelty. Despite its relatively small scale, Malta offers invaluable insights applicable to both similar and broader media contexts. These two projects focus on media monitoring and present tools designed to analyse potential biases in news articles and television news segments. The first project uses Computer Vision and Natural Language Processing techniques to analyse the coherence between images in news articles and their corresponding captions, headlines, and article bodies. The second project employs computer vision techniques to track individuals' on-screen time or visual exposure in news videos, providing queryable data. These initiatives aim to contribute to society by providing both journalists and the public with the means to identify biases. Furthermore, we make these tools accessible to journalists to improve the trustworthiness of media outlets by offering robust tools for detecting and reducing bias.

Index Terms—Artificial Intelligence (AI), AI for Journalism, Bias Detection Algorithms, Societal Implications of AI, Media Content Analysis.

I. INTRODUCTION

The world's social challenges, such as the influence of public perspectives, are being shaped as technical challenges. In this context, the societal implications of Artificial Intelligence (AI) must be given the needed weight. Therefore, society needs to advance its ethical AI practices, as outlined in the recent Council of Europe guidelines on AI in journalism [1]. The induction of bias in content is constantly challenging elections, public opinion, and democracy itself. Journalism and the media are responsible for informing the public with fair information. Information consumers can use AI to detect potential bias while also providing insights to journalists on the content they produce. As the influence of media expands globally, developing evidence-based techniques to promote fairness and accuracy is critical.

This paper presents two technical case studies on Malta's media market, which, despite being small, provide unique insights that can be extrapolated to similar contexts or broader

landscapes. These two projects explore the area of media analysis and provide tools to study potential bias in news articles and news segment videos, respectively. In the first project, we demonstrate how a range of AI techniques can be used to measure the match between the content of an image in a news article with its caption, article headline and body. In the second project, we demonstrate how computer vision techniques can be used to monitor the camera time of individuals appearing in news videos while providing queriable results. These projects directly impact society by empowering journalists and the public with tools to discern biases and inaccuracies in news content. Moreover, we aim to improve media outlets' credibility by providing tools to detect and reduce bias.

II. BACKGROUND AND LITERATURE REVIEW

This section discusses different types of media bias and its detection. Then, we present Vision Language Pre-training and Video Analysis techniques used in our case studies.

A. AI for Media Bias Mitigation

Media bias can be classified into different types [2]. Distortion bias occurs when facts are omitted, events are ignored, or numbers are presented as official without proper sourcing. Filtering bias involves cherry-picking facts when summarising an event to slant the news in a particular direction. Ideological bias promotes a specific political party or ideology, while spin bias aims to create attention-grabbing stories. Gatekeeping bias involves deciding which stories to report or suppress, while coverage bias entails unequal visibility given to certain topics. Lastly, statement bias influences reader sentiment through the language used. Contribution towards these biases can all be committed via clever usage of images in the article, namely the choice of image, position, size, visibility, description, and content.

Moreover, different automatic techniques exist to analyse these different biases [3]. For example, plagiarism detection methods can be used to identify content reuse across newspapers. News Sentiment Classifiers datasets [4] can also be applied to news article text. Facial Detection and Recognition help detect how much exposure is given to certain persons in news broadcasts or news articles. However, given the early

days of VLP technology [5], it appears that no other study leveraged the ability of VLP models to provide a multimodal bias analysis for news content, such as the approaches presented in this paper.

B. Vision Language Pre-training

Visual Language Pre-training (VLP) combines elements from CV and NLP for downstream uni-modal tasks while minimising the need for training a new model from the ground up [6]. VLP models aim to understand images in relation to natural language and develop a universal abstract understanding of how humans tend to describe pictures using words. With advancements in natural language model pre-training, namely BERT [7], transformers have become the standard architecture powering all state-of-the-art VLP models. Some examples of these large-scale VLP models are ALIGN [8] and BLIP [9]. VLP is helpful in various applications, including Caption Generation, Image-Text Retrieval, Image Classification, and Natural Language Visual Reasoning (NLVR). VLP models generally comprise three modules: the Vision Encoder, the Text Encoder, and the Multimodal Fusion. The Vision Encoder aims to extract significant features from an input image, categorised into Region Features, Grid Features, and Patch Projection. The OD-based Region Features (OD-RF) [10] approach uses an object detection model to extract bounding-box features. A drawback of OD-RF is that the vision backbone is frozen during training, which restricts overall performance. On the other hand, CNN-based Grid Features (CNN-GFs) [11] enable end-to-end (e2e) processing. Lastly, Vision Transformers (ViT) [12] split images into patches and project them into embeddings. Text encoders such as Word2Vec [13], GloVe [14] aim to create highly informative latent text representations. Due to the success of BERT [7], state-of-the-art text encoders generally employ transformers for text representation. In the multifusion module, both embeddings are combined to produce a joint representation of the image-text pair, typically done using attention. There are two types of multimodal fusion: merged attention (used in VisualBERT [15], UNITER [16], ViLT [17], and GIT [18]) and co-attention (used in LXMERT [19] and ViLBERT [10]). Merged attention concatenates features and passes them to the same transformer block, while co-attention feeds features into separate transformer blocks with cross-attention.

C. Media Bias Analysis in Videos

In their paper, Lisena et al. [20] proposed a system for face recognition in videos using a combination of MTCNN [21], FaceNet [22], and SVM classifiers. The system obtained images through crawlers, extracted face embeddings using MTCNN and FaceNet, and used an SVM classifier to identify known faces with confidence scores. Frame-level analysis was performed, with face detection, cropping, alignment, and recognition for each frame. The SORT algorithm was used for face tracking, and a weighted average algorithm was employed to label faces in frame sequences. Hierarchical clustering was used to group similar face encodings, and a ground

truth dataset was created for evaluation using news videos. Although the system demonstrates good performance, upon further investigation, it was noticed that a limited number of individuals were used within the face recognition system, which is important to consider since the error rate might increase with the number of individuals. Such systems can also be further optimised using Saliency Ranking techniques [23] that can guide the attention towards more salient regions of the scene.

Optical Character Recognition (OCR) is a process used to extract text from images, involving various steps such as pre-processing, segmentation, feature extraction, classification, and post-processing techniques such as NLP and dictionary-based approaches [24]. Challenges in OCR include dealing with different fonts or handwriting, text-background contrast, and indistinguishable characters, which can be addressed through consistent font usage and appropriate preprocessing methods [25]. Several studies have utilised OCR in analysing news videos, with Tesseract [26] being a powerful open-source OCR tool used in multiple studies [27], [28]. Wattanarachothai et al. [28] presented a novel approach for video content retrieval using text-based techniques involving key frame extraction, text localization, and keyword matching. Their method utilised Maximally Stable Extremal Region (MSER) features for key frame extraction, clustering MSERs for text localisation, and Tesseract OCR engine for text recognition. Experimental results showed the effectiveness of the MSER feature in video segmentation and highlighted improved precision and recall compared to alternative methods.

III. PROPOSED AI IN JOURNALISM TOOLS

This section presents two novel AI systems designed to help consumers better understand news content while serving as a tool for journalists.

A. News Articles Analysis

The news article analysis system uses data from six online newspapers: The Times of Malta (ToM), The Shift (TS), Malta Today (MT), The Malta Independent (MI), Malta Daily (MD), and Gozo News (GN).

The dataset includes 800 news articles from each online newspaper from the 5th of April, 2023. All articles from ToM included at least one caption describing an image. 1.75% of MT articles (14) did not have at least one caption, while this percentage was 51.00% (408) for TS. The articles from the other newspapers did not include a caption with the accompanying images. The title, article body, and image-caption pairs of every article are extracted to local storage, where six main transformations will be performed on the data: Named Entity Recognition (NER) Tagging, Keyword Extraction, Sentiment Analysis, Caption Generation, Synthetic Caption Similarity, and Image-Text Matching (ITM). Caption Generation and ITM are performed using BLIP with a ViT-L/16 [29] vision backbone. These six transformations aim to extract bias-indicative insights from the news article data, which can be

analysed and visualised by media bias researchers. Figure 1 presents the high-level architecture for this system.
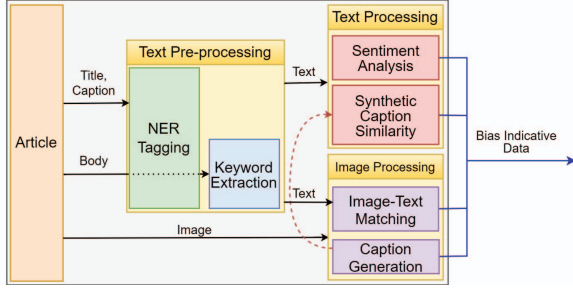

Fig. 1. Diagram showing the data transformation pipeline of the news articles data.

First, NER tagging was performed on the raw text. A list of Maltese town names was acquired from Wikipedia to improve the NER tagger's performance. Any mention of the town names in the articles was replaced with "Tokyo", as the NER tagger sometimes would not correctly label them as a location, perhaps due to their low frequency, if any, in the NER tagger's training data. Then, keyword extraction was performed on the body of the NER-tagged article. In total, 20 keywords and five single-or-double-worded keywords were extracted from the body. The Target-dependent Sentiment Classifier (TSC) model utilised [4] calculated the sentiment score for any mention of the NER tags "Person", "Organisation", and "Location" by looking at the previous and next words in the sentence. The positive, neutral, and negative scores are each averaged per article to produce an aggregate value for the current field (title, caption, or body). If none of the NER tags are present in the text, TSC is performed on the entire text with zero leading or trailing tokens. The images of the articles are passed to the VLP model BLIP [9] for caption generation. First, the images are resized to 640×640px and normalised using the default mean values of the original study. The captions are chosen using nucleus sampling [30] with the sum probability of chosen tokens equal to 0.9. The length of the generated caption is between 5 and 20 words. The Python package *sentence-transformer 2.2.2* [31] makes use of the *all-MiniLM-L6-v2* pre-trained model to calculate a similarity score between the generated caption and the NER-tagged title, caption, and body keywords. Lastly, Image-Text Matching (ITM) is performed using BLIP. The NER-tagged title, caption, and body keywords are matched to the accompanying images of the article and two similarity scores (softmax output and cosine similarity) are extracted per match.

B. News Video Analysis

This section presents an automated system for monitoring news videos using computer vision techniques. The technical contributions are as follows. Firstly, we curated and annotated a new dataset of 215 news segment videos, comprising 8.48 hours, from TVM news containing named persons in captions.

Third-party annotators carefully annotated the dataset, providing accurate timestamps, names, and flags indicating name visibility and presenter status. These annotations allow for a detailed analysis of the system's results and performance regarding OCR and person detection. The dataset was used to train and evaluate the system. The resolution of the videos ranged up to 1280×720px.

A pipeline for automatically extracting depicted persons' names using optical character recognition on caption frames was designed and built. This system was also built with privacy by design. For this reason, the system only extracts information that appears on the news video without access to external systems or sources. Moreover, it is built to be tuned to only store information about individuals holding public office to minimise any harmful application. Nonetheless, the capability for automated naming is novel. Building on this, we present our machine-learning approach for detecting, tracking and temporally localising individuals throughout the videos. This facilitates the automated generation of camera-time statistics.

The Video Analysis system comprises sub-components: face detection, recognition, object tracking, and name extraction. These work together to extract faces, track individuals, and retrieve names to create a timeline. Figure 2 shows a selection of frames from the auxiliary output provided by the system, which is intended to provide transparency and accountability when supporting the timeline illustrated in Figure 3.

1) Name Extraction: The name extraction component is tailored to extract names from videos with a specific structure. The system utilises pre-processing operations, including image closing, opening, and HSV colour thresholding. Box-like contours are identified through conditional statements and extensive testing. The preprocessed image is then processed using Pytesseract[1] OCR, which is optimised to recognise structured text, including the white box with blue text found in the videos. The most common non-empty text in a scene is selected as the individual's name to handle text transition animations. The original name is used if a match is found in the database; otherwise, the extracted name is saved in a local database. This approach successfully extracted names from most name captions in the videos.

2) Face Detection, Encoding, and Recognition: The distance between the face and the database encodings is calculated for each frame to match a detected face with a known individual. The average distance across all frames is compared to a given threshold in an adaptation of FaceRec [20]. However, an issue arises when adding an individual's facial encoding to the database due to multiple face encodings extracted from multiple frames. Different approaches were employed to select the most suitable facial encoding, including taking the first and middle frames. This was based on the work inspired by Gao et al.'s work on scene detection [32] while calculating the average of the extracted encodings. These strategies contributed to improving facial recognition accuracy.

---

[1]The Tesseract Library is available from https://pypi.org/project/pytesseract/

3) Face Tracking: Initially, attempts at scene detection were employed to segment videos into shots, but they were found to be unreliable due to sudden movements and prolonged shots that featured multiple people or individuals for only a brief moment. As a result, a tracking approach [20] was adopted to improve accuracy. The Kernelised Correlation Filters (KCF) [33] algorithm was used for scene segmentation. The KCF algorithm initialises a bounding box around a detected face and tracks it across subsequent frames. The tracker automatically adapts to changes in scale and position, ensuring the bounding box stays aligned with the tracked face. The tracker was exclusively used to indicate the start and end scenes of the individual used to generate the timeline visualisation presented in Figure 3.

Fig. 2. A frame timeline showing three people who appeared, along with the real-time analysis annotations. The generated analysis timeline can be seen in the leftmost timeline in Figure 3.

IV. EVALUATION

A. Evaluating the News Article Analysis

This section presents the evaluation of the news article analysis system. The machine-learning data-transformation pipeline was evaluated on a total of 4,800 articles from six Maltese newspapers listed in Section III-A. One of the goals of the evaluation was to examine whether newspapers artificially add to an article's perceived sentiment by intentionally using irrelevant images. The correlation coefficients in Table I show whether opinionated newspapers tend to have poorer image relevancy to the article body or vice-versa. A positive value is related to increased image relevancy as the article has more positive/neutral/negative scores. A negative value means image relevancy decreases. From the results obtained, no correlation was found between these two variables.

Another goal of the evaluation was to investigate the relationship between the images and article text. Table II summarises the main findings resulting from this experiment. The newspaper with the highest similarity score to the synthetic captions (SSC) was the Times of Malta (ToM). Malta Daily (MD) performed the best regarding Image-Text Matching (SI).

Fig. 3. The analysis timeline showing occuring individuals by their names inclduing their timestamps and durations.

TABLE I
CORRELATION COEFFICIENTS BETWEEN SENTIMENT AND IMAGE SIMILARITY FOR DIFFERENT NEWSPAPERS' ARTICLES' BODY.

Newspaper	Positive	Neutral	Negative
ToM	0.12	0.02	-0.12
TS	0.04	0.02	-0.05
MT	0.15	-0.05	-0.09
MI	-0.09	-0.02	-0.09
MD	0.08	-0.02	-0.1
GN	0.09	-0.07	-0.04

It was also noted that Gozo News (GN) articles have the least negative sentiment overall, while The Shift (TS) ranked one of the lowest in almost all metrics.

TABLE II
A TABLE OF THE AGGREGATED MEAN RESULTS OF THE DEMONSTRATION. ALL VALUES REPRESENT A PERCENTAGE.

Newspaper	SSC	SI	Positive Sent.	Neutral Sent.	Negative Sent.
ToM	**15.99**	30.99	13.87	65.55	20.58
TS	**9.27**	20.78	11.71	62.68	25.60
MT	10.80	24.15	16.79	67.03	16.18
MI	10.96	29.92	21.27	62.56	16.17
MD	13.46	**50.79**	**28.81**	58.72	12.47
GN	13.58	38.76	25.17	**69.46**	**5.37**

B. Evaluating the News Video Analysis

This section evaluates the system for automatic analysis of news videos, focusing on its strengths, weaknesses, and performance. The Video Analysis component, which forms the system's core, is essential to evaluate. While previous studies have explored news videos [27], [32], [34]–[37] and

Variation	Name extraction				Face recognition				MAE	Time
metrics	P	R	F1	A	P	R	F1	A	(secs)	(hrs)
Scene Detection										
Skips (1, 2)	0.78	0.67	0.72	0.56	0.81	0.65	0.72	0.57	10.20	4.71
Skips (2, 1)	0.81	0.67	0.73	0.58	0.84	0.65	0.73	0.57	05.73	4.27
Skips (2, 2)	0.79	0.66	0.72	0.56	0.82	0.65	0.72	0.56	09.49	2.95
Skips (2, 3)	0.81	0.67	0.73	0.58	0.82	0.63	0.71	0.56	08.45	2.42
Skips (3, 2)	0.81	0.62	0.70	0.54	0.82	0.60	0.70	0.53	05.93	1.94
Def. (2, 2)	**0.83**	**0.72**	**0.77**	0.62	0.84	0.69	0.76	0.61	09.87	4.51
Face Tracking										
Fir. (2)	0.81	0.66	0.73	0.57	0.83	0.65	0.73	0.57	07.40	2.46
Mid. (0.5)	0.78	**0.72**	0.75	0.60	0.82	0.68	0.74	0.59	**04.22**	7.59
Mid. (2)	0.81	0.68	0.74	0.59	0.83	0.63	0.72	0.56	06.89	1.98
Avg. (0.5)	0.78	0.70	0.74	0.58	0.80	0.66	0.72	0.57	04.74	7.20
Avg. (1)	0.80	0.70	0.75	0.60	0.83	0.68	0.75	0.60	05.72	4.48
Avg. (2)	0.81	0.66	0.72	0.57	0.83	0.61	0.70	0.54	07.18	**1.76**
Def. (1)	**0.83**	**0.72**	**0.77**	**0.83**	**0.86**	**0.71**	**0.78**	**0.63**	05.47	5.91
Def. (2)	0.82	0.70	0.75	0.60	0.85	0.67	0.75	0.60	07.03	4.79

face recognition in videos [20], [38], few have addressed live face recognition and name extraction in news videos simultaneously. Therefore, each sub-component will be separately assessed to determine the system's overall performance.

A comprehensive selection of metrics was employed to assess each sub-component, including name extraction, face recognition, facial tracking, and analysis duration, to evaluate the system's performance. Testing was conducted to identify optimal parameters and methods, and the results are presented in Table III, which provides an overview of the achieved performance with different variations. An incremental database test was also conducted to assess the system's scalability. Evaluating these metrics offers insights into the system's performance and areas for improvement.

1) Metrics: The performance of the name extraction subcomponent is evaluated using a confusion matrix, comparing the system's identified names against the names in the validation set per video. Using the confusion matrices, metrics such as precision, recall, F1 score, and accuracy score are calculated to assess both the name extraction and face recognition performances. To evaluate the accuracy of the scene prediction sub-component or tracker for each person, the Mean Absolute Error (MAE) is calculated. The MAE duration represents the average difference in seconds between the predicted durations and the actual durations for each individual from the annotated dataset. A smaller MAE duration indicates better predictions.

2) Results: Table III presents the performance of the Video Analysis system for the tested variations. The table is divided into two sections: one for scene detection methods and the other for the latest method using face tracking. The numbers in parentheses indicate the seconds skipped between frames during analysis. For scene detection methods, two values are displayed: the seconds skipped during analysis and the seconds skipped during scene detection. 'Skips' indicates the use of the same default parameters. 'Def.' signifies the default resolution, while 640×360px was used to speed up frame analysis in other cases. 'Fir.', 'Mid.', and 'Avg.' correspond to the first,

middle, and average face encodings, respectively, as explained in Section III-B2. In scene detection, only the first encountered face encoding was used for each individual. The metrics P (Precision), R (Recall), F1 (F1-score), and A (Accuracy), MAE (Mean Absolute Error) were calculated as described in Section IV-B1. The Time column represents the hours required to analyse all the 'News Video Segments' videos. Multiprocessing was employed in some analyses to reduce processing time. The best metric achieved for each calculation is highlighted in bold and underlined.

During the investigation, the name extraction process encountered several issues: detecting names with punctuation marks, misidentifying the right side of the caption box as 'L,' misinterpreting 'L' as the end of the caption box, exceeding frame view limitations, and skipping frames with names during video analysis. Addressing these issues is crucial to enhancing extraction effectiveness. The results show that a one-second interval for video analysis achieved the best scene detection and face tracking performance, consistent with [20]. Lower intervals in scene detection improved accuracy, while larger intervals in face tracking led to higher errors. When compared to Scene Detection and Face Tracking, Face Tracking achieved a lower MAE value, indicating better results in selecting individual timestamps. The performance of face recognition does not vary significantly with different face encoding selections. As suggested in [32], the middle face encoding generally yields slightly better results. However, relying solely on the middle or first face encoding might return lower accuracy if the face is unclear or poorly represented. The results in Table III align with existing work [32] [20] where the optimal configuration for the Video Analysis system involves using face tracking and analysing frames at 1-second intervals. The default resolutions performed slightly better, but the longer processing time does not justify this marginal advantage. Using a resolution of 640×360px is recommended for faster analysis. Overall, the results demonstrate promising

performance with potential areas for improvement in each sub-component.

V. CONCLUSION

This paper presented two technical case studies of applied AI in journalism and media, each technique presenting its novelty. We believe transparently surfacing potential distortions in news content allows media professionals of good faith opportunities to focus their efforts accordingly towards fairness, in line with the Council of Europe's recommendations on disclosure of AI use [1]. We plan for future work to expand the reach of these projects and validate them in broader contexts.

REFERENCES

[1] Council of Europe, *Guidelines on the responsible implementation of artificial intelligence systems in journalism.* adopted by the CDMSI, 2023.

[2] R. Puglisi and J. M. Snyder Jr, "Empirical studies of media bias," in *Handbook of media economics.* Elsevier, 2015, vol. 1, pp. 647–667.

[3] F. Hamborg, K. Donnay, and B. Gipp, "Automated identification of media bias in news articles: an interdisciplinary literature review," *International Journal on Digital Libraries*, vol. 20, no. 4, pp. 391–415, 2019.

[4] F. Hamborg and K. Donnay, "Newsmtsc: (multi-)target-dependent sentiment classification in news articles," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)*, Apr. 2021.

[5] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, and J. Gao, "Vision-language pretraining: Basics, recent advances, and future trends," *Foundations and Trends® in Computer Graphics and Vision*, vol. 14, no. 3–4, pp. 163–352, 2022. [Online]. Available: http://dx.doi.org/10.1561/0600000105

[6] F.-L. Chen, D.-Z. Zhang, M.-L. Han, X.-Y. Chen, J. Shi, S. Xu, and B. Xu, "Vlp: A survey on vision-language pre-training," *Machine Intelligence Research*, vol. 20, no. 1, pp. 38–56, 2023.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[8] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning.* PMLR, 2021, pp. 4904–4916.

[9] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," *arXiv preprint arXiv:2201.12086*, 2022.

[10] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[11] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," *arXiv preprint arXiv:2108.10904*, 2021.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[14] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[15] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[16] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision.* Springer, 2020, pp. 104–120.

[17] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning.* PMLR, 2021, pp. 5583–5594.

[18] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "Git: A generative image-to-text transformer for vision and language," *arXiv preprint arXiv:2205.14100*, 2022.

[19] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.

[20] P. Lisena, J. Laaksonen, and R. Troncy, "Facerec: an interactive framework for face recognition in video archives," in *DataTV 2021, 2nd International Workshop on Data-driven Personalisation of Television*, 2021.

[21] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[22] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[23] D. Seychell and C. J. Debono, "Ranking regions of visual saliency in rgb-d content," in *2018 International Conference on 3D Immersion (IC3D)*, 2018.

[24] N. Islam, Z. Islam, and N. Noor, "A survey on optical character recognition system," *arXiv preprint arXiv:1710.05703*, 2017.

[25] R. Mithe, S. Indalkar, and N. Divekar, "Optical character recognition," *International journal of recent technology and engineering (IJRTE)*, vol. 2, no. 1, pp. 72–75, 2013.

[26] Tesseract-Ocr, "Tesseract-ocr/tesseract: Tesseract open source ocr engine (main repository)." [Online]. Available: https://github.com/tesseract-ocr/tesseract

[27] R. Chatterjee and A. Mondal, "Effects of different filters on text extractions from videos using tesseract."

[28] W. Wattanarachothai and K. Patanukhom, "Key frame extraction for text based video retrieval using maximally stable extremal regions," in *2015 1st International Conference on Industrial Networks and Intelligent Systems (INISCom).* IEEE, 2015, pp. 29–37.

[29] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," 2020.

[30] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," *arXiv preprint arXiv:1904.09751*, 2019.

[31] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[32] X. Gao and X. Tang, "Unsupervised video-shot segmentation and model-free anchorperson detection for news video story parsing," *IEEE Transactions on circuits and systems for video technology*, vol. 12, no. 9, pp. 765–776, 2002.

[33] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 3, pp. 583–596, 2014.

[34] 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.

[35] W. Qi, L. Gu, H. Jiang, X.-R. Chen, and H.-J. Zhang, "Integrating visual, audio and text analysis for news video," in *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, vol. 3. IEEE, 2000, pp. 520–523.

[36] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith, "Video ocr for digital news archive," in *Proceedings 1998 IEEE International Workshop on Content-Based Access of Image and Video Database.* IEEE, 1998, pp. 52–60.

[37] J. Yang and A. G. Hauptmann, "Naming every individual in news video monologues," in *Proceedings of the 12th annual ACM international conference on Multimedia*, 2004, pp. 580–587.

[38] H. Wang, Y. Wang, and Y. Cao, "Video-based face recognition: A survey," *International Journal of Computer and Information Engineering*, vol. 3, no. 12, pp. 2809–2818, 2009.