World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Mouth detector with combination between LipNet and PyImageSearch

Le Minh Quan Tran [1,*] and Linh Nguyen Hoang Anh [2]

[1] *Mechatronic Systems Engineering, Faculty of Technology and Bionics, Hochschule Rhein-Waal, Germany.*
[2] *Information Technology, Faculty of Interdisciplinary Science, Vietnam National University Ho Chi Minh City - University of Science, Vietnam.*

World Journal of Advanced Research and Reviews, 2024, 22(03), 065–073

Publication history: Received on 17 April 2024; revised on 29 May 2024; accepted on 01 June 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.22.3.1633

Abstract

In the digital age, the authenticity and integrity of multimedia content have become increasingly critical, especially in contexts involving high-profile political figures like former President Donald Trump. The proliferation of manipulated and deepfake videos poses significant threats to public discourse and trust. This project addresses these challenges by developing a lip-reading system using PyImageSearch and Python TensorFlow to verify and validate spoken content in videos. By leveraging advanced computer vision techniques and machine learning models, the system aims to analyze and interpret lip movements accurately, ensuring the correspondence between the spoken words and the audio track. This tool can be instrumental in detecting inconsistencies and potential manipulations in videos, thereby enhancing the reliability of digital content. The ultimate goal is to provide a robust solution that can be utilized by media organizations, fact-checkers, and the general public to maintain the integrity of political communications in the digital realm.

Keywords: Mouth Open Detector; Facial Recognition; PyImage Search; LipNet

1. Introduction

In today's increasingly digitized world, the accurate interpretation of human gestures and expressions is crucial for enhancing user experiences across diverse domains. Among these gestures, the act of mouth opening holds particular significance, serving as a fundamental indicator of both physiological and emotional states.

In light of contemporary advancements in artificial intelligence (AI) technology, particularly in the realm of video manipulation, there arises a significant concern regarding the authenticity and integrity of digital content, particularly in instances where political figures such as former President Donald Trump are involved. Specifically, the propagation of videos purporting to capture speeches or statements by prominent individuals, such as Trump's address at the World Government Conference, demands careful scrutiny. Malicious entities within society have been known to employ advanced AI technologies with the intention of impugning honor and dignity. By manipulating video content to synchronize Trump's voice with his oral articulations, these bad actors aim to create fabricated media that could potentially instigate significant political controversies. The propagation of such manipulated media may result in misunderstandings and the exacerbation of conflicts between nations. The potential consequences of this misuse of AI technology are far-reaching and could undermine the integrity of political discourse, international relations, and social cohesion.

Given the gravity of this issue, it is imperative to develop robust countermeasures to prevent the malicious manipulation of digital media with urgency. In this proposal, I advocate for the development of a Mouth Open Detector (MOD), a technological innovation poised to significantly impact various sectors by leveraging advanced computer vision

* Corresponding author: Le Minh Quan Tran

techniques and machine learning algorithms. The technology enables users to detect the subject's mouth shape, analyze its movements and verify if the subject's voice aligns with those movements.

2. Research Objectives

The primary objective of this research is to develop and implement an effective MOD to address the following key points:

Figure 1 Illustration of how shape of speaking a certain characters

- Detection of Manipulated Media: MOD employs advanced computer vision techniques and cutting-edge machine learning algorithms to precisely identify situations in which video content has been manipulated to synchronize a subject's speech with artificial lips movements. The system detects potential cases of digital forgeries through the analysis of differences between audio and visual cues.
- Verification of Authenticity: Through comprehensive analysis of moth motions and forms in relation to the accompanying audio, this technology offers a method for verifying the legitimacy of digital media material. MOD determines whether the content is genuine or manipulated by confirming the alignment between human's spoken words and their corresponding oral articulations.
- Prevention of Misinformation: This state-of-the-art development is essential in preventing the spread of misinformation and disinformation by efficiently identifying and reporting distorted media. By empowering users to distinguish between authentic and fabricated content, the technology potentially lessens the detrimental effects of false narratives and political manipulation on public discourse and societal trust.
- Enhancement of Trust in Digital Media: MOD contributes to improve confidence in media sources and information dissemination platforms by promoting authenticity and transparency in digital content. In an era characterized by pervasive misinformation, the technology fosters increased trust in the reliability and credibility of digital media by providing people the capacity to independently check the integrity of video material.
- Protection of Political Integrity and Social Cohesion: By safeguarding against the harmful alteration of films featuring political figures, including former President Donald Trump, MOD contributes to maintain the integrity of international relations and political debate. This implementation promotes greater stability and cohesiveness throughout society by minimizing the potential that fabricated media would ignite political disputes and exacerbate social divisions.

Overall, the development and implementation of MOD represent a vital step in mitigating the hazards connected to the improper application of AI technology in the manipulation of digital media. The technology has the potential to significantly impact various sectors, like politics, media and social discourse, by enabling the accurate detection and verifying distorted material, ultimately contributing to a more unified, trustworthy and cohesive society.

3. Related Works

The development of the MOD is situated within a dynamic field of research focusing on facial recognition and mouth detection technologies. Several studies provide foundational insights that inform and contextualize this project:

The first study titled "An Efficient Mouth Detection Based on Face Localization and Edge Projection"[1] presents a method for detecting the mouth within an image by first isolating the face and then employing edge projection techniques to accurately delineate the mouth area. The methodology involves segmenting the face using a combination of skin color detection and geometrical properties, followed by the application of edge detection algorithms specifically focused on the lower facial region to pinpoint the mouth's boundaries. This technique is particularly advantageous in applications such as speech recognition systems, where precise mouth detection can significantly enhance the system's ability to interpret speech accurately. It is also beneficial in security and identification systems where verifying individuals' identities through their facial features is crucial. By isolating the mouth accurately, the technology improves the performance of facial recognition systems, enhancing their reliability and effectiveness in controlled environments.

Despite its strengths, this method exhibits certain limitations when applied in less controlled or variable conditions. The primary weaknesses include:

- Dependency on Lighting and Resolution: The technique's effectiveness is heavily dependent on good lighting conditions and high image resolution. In low-light or uneven lighting conditions, the edge detection component struggles to accurately identify the mouth's boundaries.
- Sensitivity to Orientation and Expression: Changes in facial orientation or expressions can significantly affect the accuracy of mouth detection. The method might not reliably detect the mouth in images where the face is not directly facing the camera or in expressions that obscure typical mouth edge features.
- Generalization Across Diverse Environments: While the method performs well in controlled settings, its adaptability to diverse and dynamic environments is limited. This restricts its utility in applications that require operation in outdoor or variable lighting conditions, or where faces are captured at different angles or in motion.

In contrast, the MOD proposed in our project seeks to address these limitations by utilizing advanced deep learning algorithms that are not as reliant on perfect lighting or resolution, and which can handle variations in facial orientation and expressions more adeptly. This makes MOD a more robust solution for real-world applications, extending its usability across a wider range of environments and scenarios.

The second study described in "The process of lips detection system"[2] outlines a comprehensive approach to lip detection which involves multiple stages: image preprocessing, lip region extraction, and color segmentation. This method initially preprocesses the image to enhance the visual clarity and define the lip region. Following this, the technique uses specific color segmentation strategies to isolate the lips from other facial features based on distinct chromatic properties. This process is instrumental in identifying the precise boundaries and features of the lips. This system is highly applicable in fields such as digital telecommunication and cosmetic testing, where accurate lip detection can significantly enhance product interfaces and user experiences. For example, in video conferencing tools, precise lip detection can improve speech clarity and synchronization, enhancing communication effectiveness. Similarly, in virtual makeup applications, accurate lip detection allows for better simulation of cosmetic products on a user's lips, providing a more realistic and satisfying user experience. While this system is effective in certain applications, it has several notable limitations:

- Dependence on Uniform Lighting Conditions: The accuracy of color segmentation heavily relies on consistent lighting. Variations in lighting can lead to incorrect lip detection as changes in shadow and highlight can alter the perceived color of the lips.
- Skin Tone Variability: The system's effectiveness can diminish across different skin tones. Since color segmentation is based on specific chromatic thresholds, discrepancies in skin tones can result in errors, making the system less versatile across a diverse user base.
- Static and Controlled Settings: The system primarily functions optimally in static settings where the subject's face is steady and frontal. It is less effective in dynamic scenarios where the subject is moving, or the facial orientation changes, which are common in real-world applications.

In contrast, the MOD seeks to overcome these challenges by integrating advanced machine learning techniques that do not solely rely on color segmentation. MOD employs a combination of phoneme-to-viseme analysis and biometric verification, enabling it to function effectively under varied lighting conditions, across diverse skin tones, and in dynamic environments. This approach not only improves the robustness of lip detection but also broadens the applicability of the technology to more real-world scenarios, making it a superior choice for applications requiring reliable and inclusive facial feature detection.

The review of existing methodologies within the facial recognition and mouth detection domains - specifically the techniques centered around facial localization with edge projection and color-segmented lip detection - underscores both the progress and the persistent limitations in this field. While these methods have proven effective in controlled environments, their real-world applicability remains constrained by several critical factors such as dependency on consistent lighting conditions, limited adaptability across diverse skin tones, and sensitivity to dynamic facial orientations and expressions.

MOD is designed to transcend these limitations, offering a robust solution that leverages advanced machine learning algorithms and a comprehensive phoneme-to-viseme analysis. This technology not only ensures accuracy in mouth detection under a variety of challenging conditions but also maintains high performance regardless of lighting variability and facial dynamics. MOD's sophisticated audio-visual synchronization analysis further enables it to operate

efficiently in diverse and uncontrolled environments, enhancing its utility across numerous practical applications. Beyond technical superiority, MOD's development carries significant societal implications. By accurately identifying and verifying mouth movements, MOD plays a crucial role in combating digital misinformation, a growing concern in today's media landscape. Its capability to discern authenticity from manipulated content bolsters the integrity of information disseminated across digital platforms, thereby fostering a more informed and less polarized public discourse.

4. Process

4.1. Overall Process

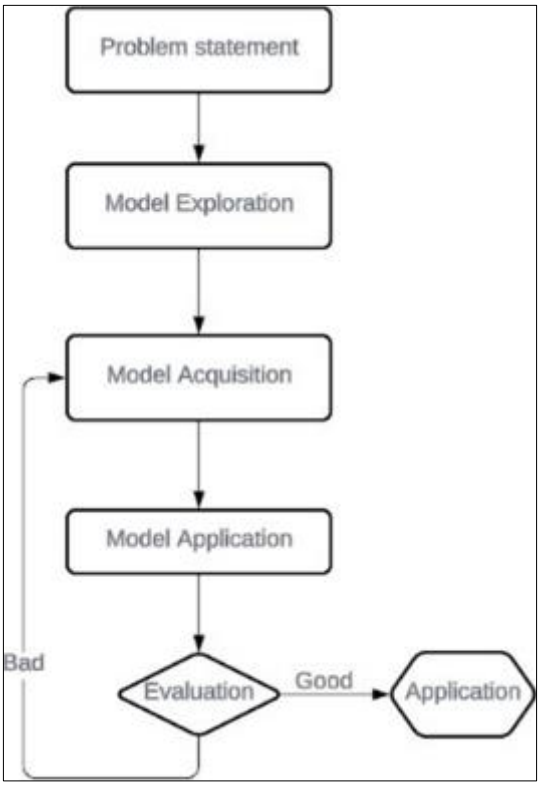

Figure 2 Illustration of overall process

To be more specific:

- Problem statement: Observe, study and identify issues throughout videos on Internet
- Model exploration: Explore existing model and determine the scope of the obtained model
- Model acquisition: Evaluate the ability of applying in project
- Model Application: Apply and build models for new approach
- Evaluation: Assess the model using evaluation metrics such as accuracy and confusion matrix.

4.2. Mouth Open Detector Algorithm

We'll start by setting up a camera to search for faces in a stream. In the event that a face is discovered, we extract the mouth areas using facial landmark detection. With the mouth regions in place, we can calculate the mouth aspect ratio [3] to ascertain whether the mouth is open or closed. We will predict what individuals are talking about using scripts if the mouth aspect ratio shows that the mouth has been open for a significant amount of time. We'll use Steamlit, OpenCV, and Python to create the Lip Read detection technique described above in the following section.

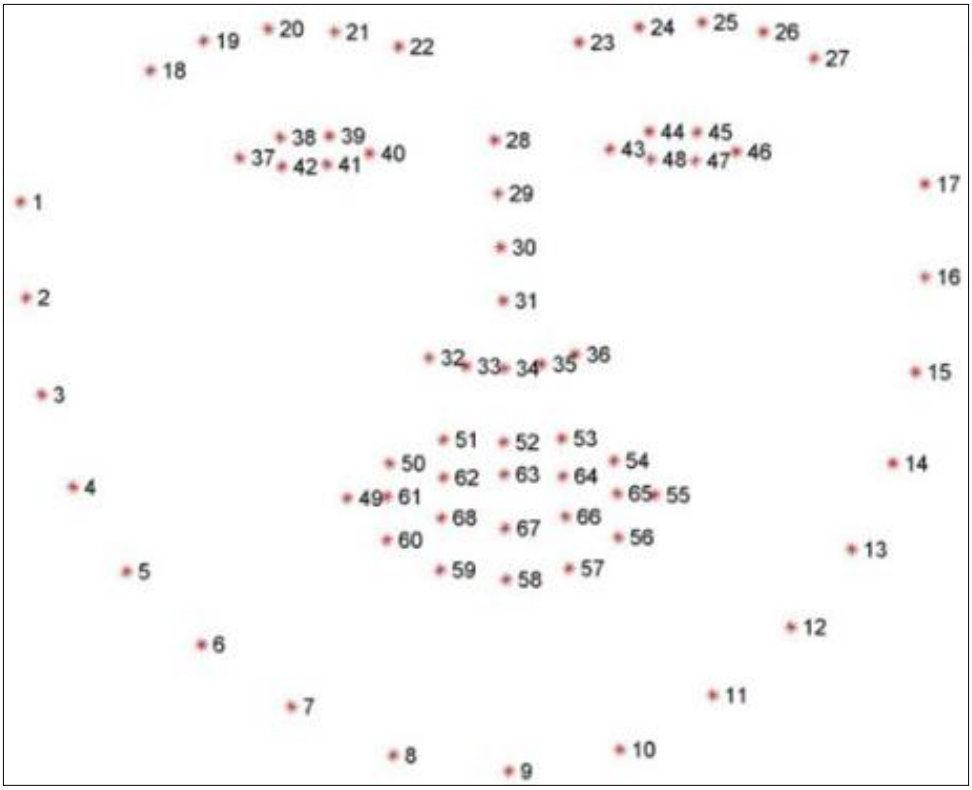

Figure 3 Visualizing the 68 facial landmark coordinates from the iBUG 300-W dataset [4]

4.3. Theory Basic

4.3.1. PyImageSearch

PyImageSearch [5] is a widely recognized platform and blog dedicated to computer vision and image processing using Python. Founded by Adrian Rosebrock, PyImageSearch has become a go-to resource for developers, researchers, and enthusiasts looking to delve into the world of computer vision. The platform offers a wealth of tutorials, code snippets, and practical guides that cover a broad spectrum of topics, from basic image processing techniques to advanced machine learning applications. One of the key strengths of PyImageSearch is its focus on making complex computer vision concepts accessible and easy to understand. It provides step-by-step instructions and hands-on projects that allow users to apply theoretical knowledge to real-world problems. The tutorials often include detailed explanations, code examples, and visual aids, ensuring that learners can follow along regardless of their prior experience level.

PyImageSearch also emphasizes the use of popular Python libraries such as OpenCV, TensorFlow, and Keras, demonstrating how these tools can be leveraged to build powerful computer vision applications. Topics covered include object detection, face recognition, image classification, and deep learning, among others. By integrating these libraries, PyImageSearch helps users to develop efficient and scalable solutions for a variety of image processing tasks.

4.3.2. LipNet

LipNet utilizes a convolutional neural network (CNN) combined with a recurrent neural network (RNN) to capture spatial and temporal features from video sequences:

- Convolutional Layers: These layers extract spatial features from individual video frames, identifying critical patterns and movements associated with lip shapes and positions.
- Recurrent Layers: Specifically, a bidirectional Long Short-Term Memory (BLSTM) network is used to model the temporal dynamics of lip movements. By processing information in both forward and backward directions, BLSTM effectively captures context and dependencies across the video frames.
- Connectionist Temporal Classification (CTC) Loss: LipNet employs CTC loss to align the predicted sequences with the ground truth labels without requiring precise frame-level annotations. This allows the model to handle varying lengths of video sequences and speech rates naturally.

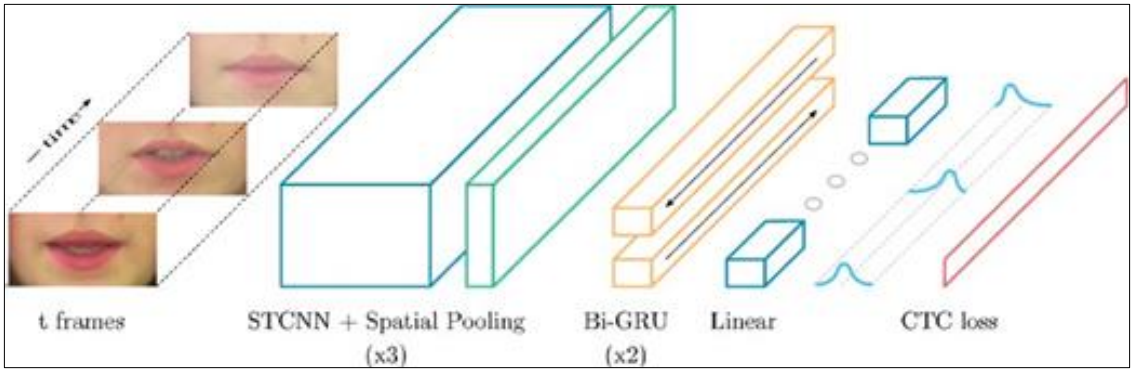

Figure 4 LipNet architecture. A sequence of T frames is used as input, and is processed by 3 layers of STCNN, each followed by a spatial max-pooling layer. [6]

5. Result

The model is trained on large datasets of video clips with corresponding transcriptions. One prominent dataset used in LipNet's development is the GRID corpus, which contains thousands of sentences spoken by multiple speakers. The training process involves optimizing the model to minimize the CTC loss, thereby improving its ability to accurately predict sequences of text from video input. LipNet's performance has set new benchmarks in the field of automatic lip reading. It has achieved higher word accuracy rates compared to previous methods, demonstrating its capability to generalize across different speakers and various speaking conditions. This advancement is particularly significant for applications in silent communication systems, hearing impairment aids, and security and surveillance.

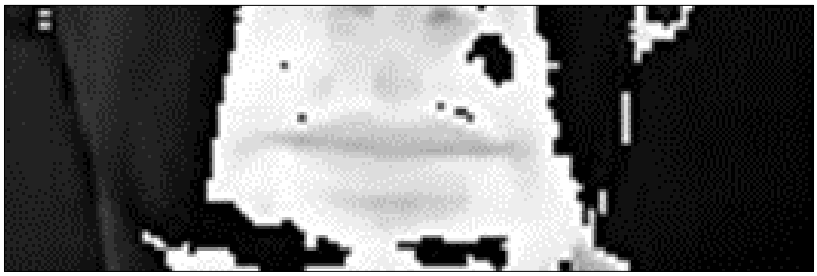

Figure 5 Application on random man talking

The speed of the model is always shorter than 5 seconds per video under 10 seconds. The subtitle for what people are talking about in the video is higher than 60% for accuracy in a total of 100 videos of testing.

LipNet's success underscores the potential of deep learning in transforming lip reading from a specialized skill to a widely accessible technology. Future research directions include: Dataset Expansion: Incorporating more diverse datasets to improve robustness across different languages, dialects, and speaking styles.

Integration with Audio: Combining lip reading with audio signals for improved speech recognition in noisy environments.

Real-time Applications: Enhancing the model's efficiency for real-time deployment in interactive systems and assistive technologies.

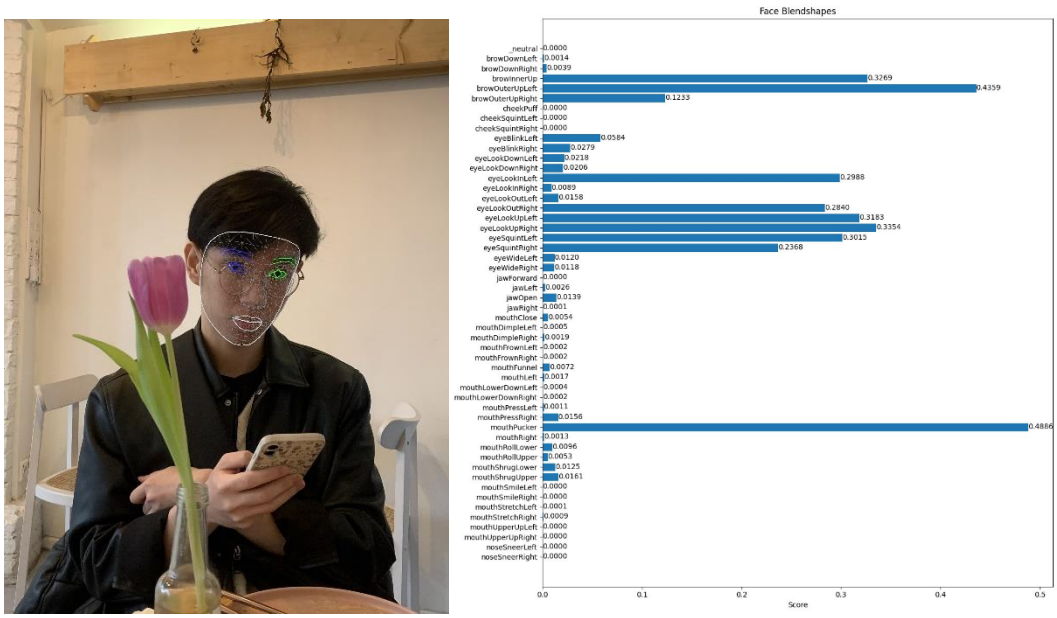

Figure 6 Result on testing images with obstructions on image

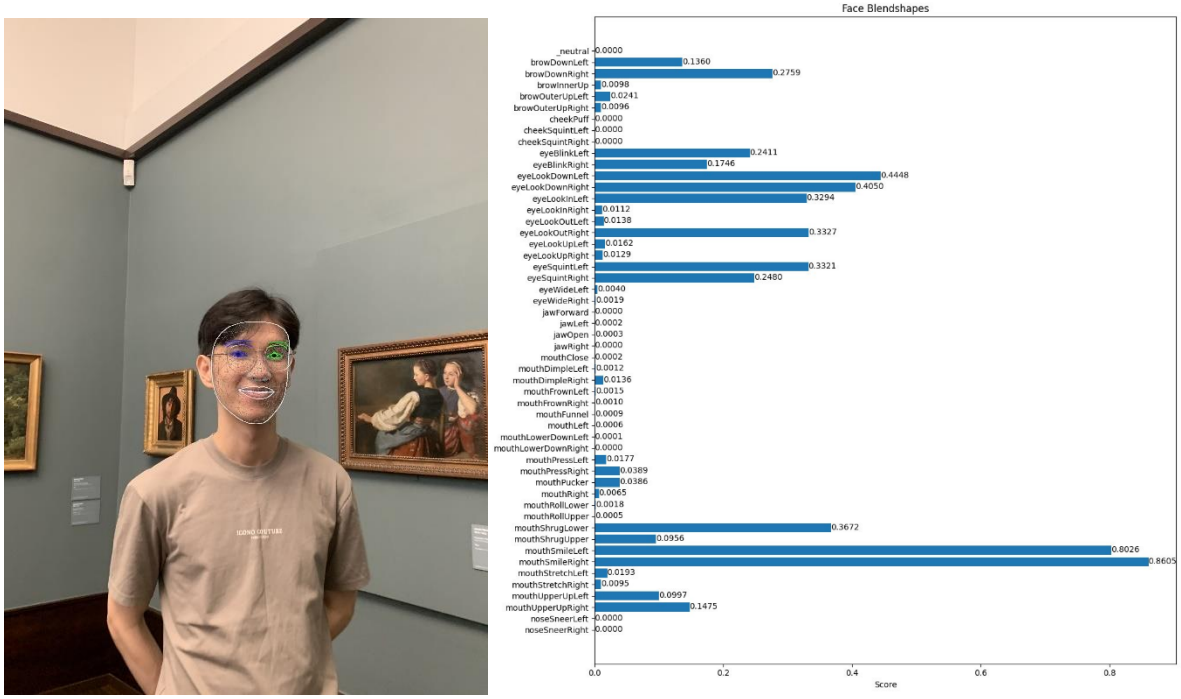

Figure 7 Result on testing images with parameters

In summary, LipNet represents a significant milestone in the quest to develop accurate and efficient lip reading systems. Its innovative use of deep learning techniques not only enhances the capabilities of automatic speech recognition but also opens new avenues for research and applications in computer vision and human-computer interaction.

6. Discussion

The development and implementation of LipNet highlight significant advancements in the field of automatic lip reading, showcasing the power of deep learning in addressing complex computer vision challenges. This discussion will explore

the implications, limitations, and future directions of LipNet, providing a comprehensive overview of its impact and potential.

6.1. Implications for Technology and Society

LipNet's ability to accurately read lips from silent video sequences has profound implications for various domains:

Accessibility: For individuals with hearing impairments, LipNet can be integrated into assistive technologies, enhancing communication by providing real-time speech-to-text translation from lip movements. This could significantly improve accessibility in noisy environments or situations where audio is unavailable.

Security and Surveillance: In security applications, LipNet can be utilized to monitor and interpret speech in video surveillance footage without relying on audio, which may be compromised or intentionally muted. This capability is valuable in scenarios where understanding spoken content is critical for security operations.

Multimedia and Entertainment: LipNet could be employed in the entertainment industry for tasks such as dubbing and improving the accuracy of subtitles in videos, especially in cases where the audio quality is poor or languages need to be translated without relying solely on audio cues.

6.2. Limitations

Despite its promising performance, LipNet has several limitations that need to be addressed:

- Generalization Across Diverse Conditions: While LipNet has shown high accuracy on datasets like the GRID corpus, its performance may vary with different languages, accents, and speaking styles. Real-world conditions often present diverse challenges, such as variations in lighting, camera angles, and background noise, which can affect lip reading accuracy.
- Dependence on High-Quality Data: The effectiveness of LipNet depends on the quality and quantity of training data. Collecting and annotating large datasets with accurate transcriptions is resource-intensive. Moreover, datasets need to be representative of real-world scenarios to ensure the model's robustness.

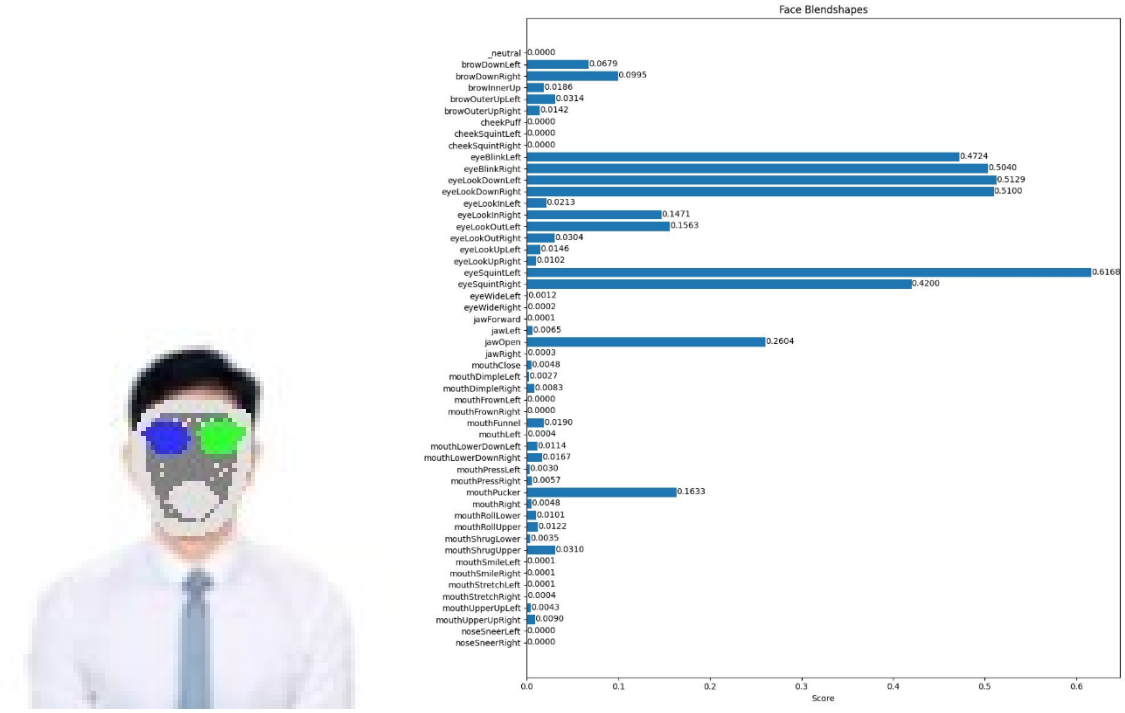

Figure 8 Illustration of a result on low-quality data

- Computational Complexity: The deep learning architecture of LipNet, particularly the combination of CNN and BLSTM layers, requires significant computational resources for training and inference. This can limit its deployment in resource-constrained environments or real-time applications without further optimization.

6.3. Future Directions

To address these limitations and expand the capabilities of LipNet, several future research directions are proposed:

- Dataset Diversification: Expanding the training datasets to include more diverse speakers, languages, and real-world conditions will enhance the model's generalization and robustness. Additionally, creating multilingual datasets could facilitate LipNet's application in different linguistic contexts.
- Integration with Multimodal Inputs: Combining lip reading with other modalities, such as audio signals and facial expressions, could improve the accuracy and reliability of speech recognition systems. Multimodal models can leverage complementary information to overcome challenges posed by visual-only or audio-only inputs.
- Model Optimization: Research into optimizing the model architecture and employing techniques such as model pruning, quantization, and edge computing can reduce the computational requirements, making LipNet more feasible for real-time applications and deployment on mobile devices.
- User-Centered Design: Developing user-friendly interfaces and applications that leverage LipNet's capabilities will ensure its practical utility. Collaborating with end-users, particularly those with hearing impairments, will provide valuable insights into the design and functionality of assistive technologies.

7. Conclusion

The development of the MOD heralds a significant leap in scientific research aimed at ensuring the authenticity of digital media, particularly in politically sensitive contexts. Through the sophisticated application of computer vision and machine learning, MOD meticulously analyzes lip movements to detect video manipulations, addressing the pervasive issue of misinformation. This pioneering technology promises to elevate the integrity and transparency of digital content, fostering trust and credibility in media communications. Ultimately, MOD stands as a critical tool for safeguarding political discourse and promoting an informed, cohesive society.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] An Efficient Mouth Detection Based on Face Localization and Edge Projection IJCTE. [Online]. Available: https://www.ijcte.org/index.php?m=content&c= index&a=show&catid=49&id=857.

[2] New Lips Detection and Tracking System. Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol IIMECS 2009, Mar. 2009.

[3] R. A., Facial landmarks with dlib, OpenCV, and Python, PyImageSearch. 2021. [Online]. Available: https://pyimagesearch.com/2017/04/03/facial- landmarks-dlib-opencv-python/.

[4] 300 Faces In-the-Wild Challenge (300-W), ICCV 2013 i·bug - resources - 300 Faces In-the-Wild Challenge (300-W), ICCV 2013. [Online]. Available: https:// ibug.doc.ic.ac.uk/resources/300-W/.

[5] You can become a computer vision + OpenCV Guru. PyImageSearch Gurus: Computer Vision and OpenCV Course. [Online]. Available: https : / / pyimagesearch.com/pyimagesearch-gurus/.

[6] Assael, Y., Shillingford, B., Whiteson, S., & Freitas, N.D. (2016). LipNet: End-to-End Sentence-level Lipreading. arXiv: Learning.