DEMOCRATIC AND POPULAR REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND RESEARCH

SAAD DAHLAB BLIDA UNIVERSITY - 1
Faculty of Science
Computer Science Department
LRDSI Laboratory

DOCTORAL THESIS OF THE THIRD CYCLE
Option: software engineering

Detection and tracking of targets in real-time images Case: Arabic Visual Speech Recognition System BlidaAVS10 (Lipreading)

Realized by:
Ali Baaloul

Supervisor:
Pr. Nadjia Benblidia

In Front of The Jury :

Oukid-Khouas Saliha	Professor, University of Blida 1	President
Merouani Hayet	Professor, University Badji Mokhtar, Annaba	Examiner
Boustia Narhimene	Professor, University of Blida 1	Examiner
Ouared Abdelkader	Associate Professor, University of Tiaret	Examiner
Nadjia Benblidia	Professor, University of Blida 1	Supervisor

Algeria, Blida, 2024

I dedicate my thesis work to my:
late father, my dear mother
loving wife and our children
supportive siblings
beloved family
All cherished friends
. . .

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Nadia Benblidia, for her constant guidance, encouragement, and feedback throughout this research project. She has been a source of inspiration and motivation for me, and I have learned a lot from her expertise and experience.

I am thankful to the members of my thesis committee, Prof. A, Dr. B, and Dr. C, for their constructive criticism and insightful questions during my defense. Their inputs have helped me improve the quality and clarity of my thesis.

I appreciate the support and collaboration of my colleagues and friends at the LRDSI laboratory and outside, who have shared their knowledge, skills, and resources with me. They have also made my research journey more enjoyable and memorable.

Last but not least, I would like to thank my family, especially my parents, my wife for their unconditional love, care, and sacrifice. They have always been there for me, no matter what. I dedicate this thesis to them....

Ali Baaloul
baaloulali@gmail.com

Abstract

0.1 English

Automatic visual speech recognition (AVSR) techniques are increasingly prevalent in various domains, including manufacturing, public use, and multimedia devices, making Visual Speech Recognition (VSR) a promising technology that can improve communication accessibility for people with hearing impairments. However, most existing VSR systems are designed for languages like English, leaving a gap for languages like Arabic, which is spoken by over 400 million people worldwide and has unique linguistic and phonetic characteristics. This thesis presents a novel framework for Arabic Visual Speech Recognition, which aims to address this gap and cater to the needs of the Arabic hearing impaired community. The framework integrates state-of-the-art deep learning techniques, such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViT), to transcribe Arabic speech from visual cues accurately and efficiently. The framework also relies on a specialized Arabic dataset, which is carefully curated to capture the diversity and complexity of the Arabic language. This dataset serves as a benchmark for training and evaluating the VSR models, ensuring their robustness and reliability in real-world applications. The framework employs the deep learning techniques like YOLO, CNNs and ViT for robust mouth detection and recognition, which enables the extraction of crucial visual features for accurate speech transcription. The experimental results show that the proposed framework achieves promising performance in enhancing communication accessibility for Arabic speakers with hearing impairments. The framework also demonstrates its effectiveness in handling various linguistic and phonetic variations of the Arabic language, opening up new possibilities for wider applications in real-world scenarios. This research contributes significantly to advancing Arabic Visual Speech Recognition technology, enriching the VSR landscape and fostering greater inclusivity in communication for Arabic speakers.

Key words:
Lip-readings, Visual Speech Recognition, Audiovisual dataset, CNN, Vision Transformer

0.2 Frensh

Les techniques de reconnaissance automatique de la parole visuelle (ARSP) sont de plus en plus répandues dans divers domaines, notamment la fabrication, l'utilisation publique et les appareils multimédias, faisant de la reconnaissance visuelle de la parole (RVP) une technologie prometteuse pouvant améliorer l'accessibilité à la communication pour les personnes malentendantes. Cependant, la plupart des systèmes RVP existants sont conçus pour des langues telles que l'anglais, laissant un vide pour des langues comme l'arabe, parlée par plus de 400 millions de personnes dans le monde et possédant des caractéristiques linguistiques et phonétiques uniques. Cette thèse présente un nouveau cadre pour la reconnaissance visuelle de la parole en arabe, qui vise à combler ce vide et à répondre aux besoins de la communauté arabe malentendante. Le cadre intègre des techniques d'apprentissage en profondeur de pointe, telles que les réseaux de neurones convolutifs (CNN) et les transformateurs de vision (ViT), pour transcrire avec précision et efficacité la parole arabe à partir de signaux visuels. Le cadre s'appuie également sur un ensemble de données arabe spécialisé, soigneusement sélectionné pour capturer la diversité et la complexité de la langue arabe. Cet ensemble de données sert de référence pour la formation et l'évaluation des modèles RVP, garantissant leur robustesse et leur fiabilité dans des applications du monde réel. Le cadre utilise des techniques d'apprentissage en profondeur comme YOLO, les CNN et ViT pour une détection et une reconnaissance robustes des bouches, ce qui permet l'extraction de caractéristiques visuelles cruciales pour une transcription de la parole précise. Les résultats expérimentaux montrent que le cadre proposé atteint des performances prometteuses pour améliorer l'accessibilité à la communication pour les locuteurs arabophones malentendants. Le cadre démontre également son efficacité dans la gestion des variations linguistiques et phonétiques de la langue arabe, ouvrant de nouvelles possibilités pour des applications plus larges dans des scénarios du monde réel. Cette recherche contribue de manière significative à faire progresser la technologie de reconnaissance visuelle de la parole en arabe, enrichissant le paysage de la RVP et favorisant une plus grande inclusivité dans la communication pour les locuteurs arabophones.

Mots clés: *Lecture labiale, Reconnaissance visuelle de la parole, Ensemble de données audiovisuelles, CNN (Réseaux de neurones convolutifs), Transformateur de vision.*

0.3 Arabic

تنتشر تقنيات التعرف التلقائي على الكلام المرئي بشكل متزايد في مجالات مختلفة، بما في ذلك التصنيع والاستخدام العام وأجهزة الوسائط المتعددة، مما يجعل التعرف على الكلام المرئي تقنية واعدة يمكنها تحسين إمكانية الوصول إلى الاتصالات للأشخاص الذين يعانون من ضعف السمع. ومع ذلك، فإن معظم أنظمة طصغ الحالية مصممة للغات مثل اللغة الإنجليزية، مما يترك فجوة للغات مثل اللغة العربية، التي يتحدث بها أكثر من ٤٠٠ مليون شخص حول العالم وتتمتع بخصائص لغوية وصوتية فريدة. تقدم هذه الأطروحة إطارًا جديدًا للتعرف البصري على الكلام العربي، والذي يهدف إلى معالجة هذه الفجوة وتلبية احتياجات مجتمع ضعاف السمع العربي. يدمج الإطار أحدث تقنيات التعلم العميق، مثل الشبكات العصبية التلافيفية ومحولات الرؤية، لنسخ الكلام العربي من الإشارات المرئية بدقة وكفاءة. ويعتمد الإطار أيضًا على مجموعة بيانات عربية متخصصة، تم تنسيقها بعناية لالتقاط تنوع اللغة العربية وتعقيدها. تعمل مجموعة البيانات هذه كمعيار للتدريب وتقييم نماذج التعرف على الكلام المرئي، مما يضمن قوتها وموثوقيتها في تطبيقات العالم الحقيقي. يستخدم الإطار تقنيات التعلم العميق مثل نوبو وث س وطة لاكتشاف الفم والتعرف عليه بشكل قوي، مما يتيح استخلاص الميزات المرئية المهمة لنسخ الكلام بدقة. أظهرت النتائج التجريبية أن الإطار المقترح يحقق أداء واعدا في تعزيز إمكانية الوصول إلى الاتصالات للناطقين باللغة العربية من ذوي الإعاقة السمعية. يُظهر الإطار أيضًا فعاليته في التعامل مع الاختلافات اللغوية والصوتية المختلفة للغة العربية، مما يفتح إمكانيات جديدة لتطبيقات أوسع في سيناريوهات العالم الحقيقي. يساهم هذا البحث بشكل كبير في تطوير تقنية التعرف على الكلام المرئي باللغة العربية، وإثراء مشهد طصغ وتعزيز المزيد من الشمولية في التواصل للمتحدثين باللغة العربية.

الكلمات المفتاحية:

قراءة الشفاه ، التعرف على الكلام المرئي، الشبكات العصبية التلافيفية، المتحولات البصرية.

Contents

List of Figures

List of Tables

List of Abbreviations

ACM	**A**ctive **C**ontour **M**odels
AAM	**A**ctive **A**ppearance **M**odels
ALR	**A**udio-**V**isual **S**peech **R**ecognition
AVSD	**A**rabic **V**isual **S**peech **D**ataset
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
CNNs	**C**onvolutional **N**eural **N**etworks
CPU	**C**entral **P**rocessing **U**nit
CV	**C**omputer **V**ision
DCT	**D**iscrete **C**osine **T**ransform
DBNF	**D**eep **B**ottleneck **N**eural **F**eatures
Dlib	**D**esigned for machine learning, computer vision, and image processing tasks
EM	**E**xpectation-**M**aximization
FC	**F**uzzy **C**lustering
FPS	**F**rames **P**er **S**econd
GPU	**G**raphics **P**rocessing **U**nit
GRU	**G**ated **R**ecurrent **U**nit
HMM	Hidden Markov Models
HOG	**H**istogram of **O**riented **G**radients
KLT	**K**anade-**L**ucas-**T**omasi
LDA	**L**inear **D**iscriminant **A**nalysis
LSTM	**L**ong **S**hort-**T**erm **M**emory
MLLT	**M**aximum **L**ikelihood **L**inear **T**ransform
MHT	**M**ultiple **H**ypothesis **T**racking
MSA	**M**odern **S**tandard **A**rabic
PCA	**P**rincipal **C**omponent **A**nalysis
R-CNN	**R**egion-based **C**onvolutional **N**eural **N**etworks
RNNs	**R**ecurrent **N**eural **N**etworks
RPN	**R**egion **P**roposal **N**etwork
RoI	**R**egion **o**f **I**nterest
RGB	**R**ed, **G**reen, and **B**lue
SIFT	**S**cale-**I**nvariant **F**eature **T**ransform
SORT	**S**imple **O**nline and **R**ealtime **T**racking
SVM	**S**upport **V**ector **M**achines
SSD	**S**ingle **S**hot **M**ultibox **D**etector
ViT	**V**ision **T**ransformers
VLRF	**V**isual **L**ip **R**eading **F**eatures
VSR	**V**isual **S**peech **R**ecognition
YOLO	**Y**ou **O**nly **L**ook **O**nce

Part I

General Introduction

Chapter 1

Introduction

1.1 Motivation

The 21st century has witnessed the transformative rise of machine learning and deep learning, enabling autonomous learning from data and revolutionizing tasks like image recognition and natural language processing. These technologies address complex challenges across diverse fields, driving innovation and improving human life. In this thesis, we harness machine learning and deep learning for visual speech recognition, showcasing their impact on human-computer interaction and accessibility. Visual Speech Recognition (VSR) emerges as a compelling application, with the potential to address societal challenges in communication and accessibility. This subsection explores the profound significance of VSR, particularly in aiding individuals with hearing impairments and dismantling communication barriers.

Empowering Individuals with Hearing Impairments One of the most compelling and transformative aspects of VSR lies in its ability to empower individuals with hearing impairments. For those who experience challenges in auditory perception, traditional speech communication may pose formidable obstacles. Visual speech recognition harnesses the complexity interplay between lip movements, facial expressions, and other visual cues to decipher spoken language. By translating these visual cues into comprehensible textual or auditory output, visual speech recognition provides a means for individuals with hearing impairments to engage in effective communication.

The potential impact of this technology on the lives of the hearing-impaired community is profound. It offers the promise of enhanced educational experiences, improved social interactions, and increased opportunities for employment. The ability to participate fully in conversations, lectures, and various communication contexts can substantially elevate the quality of life for individuals with hearing impairments.

Overcoming Communication Barriers Visual speech recognition also holds immense promise in transcending communication barriers that arise in multilingual and multicultural settings. In scenarios where individuals speak different languages or dialects, visual cues can often serve as a universal medium for conveying information. By enabling real-time translation of spoken language into written text or other forms of communication, visual speech recognition fosters cross-cultural understanding and collaboration.

Moreover, visual speech recognition has the potential to bridge the gap between spoken and sign languages. While sign language is a critical mode of communication for many individuals with hearing impairments, there remains a disparity in its widespread adoption and understanding. Visual speech recognition can serve as an intermediary, facilitating

communication between those who use sign language and those who rely on spoken language. This inclusive approach has far-reaching implications for creating more inclusive and accessible communication ecosystems.

Technological Advancements and Inclusive Communication The cross-disciplinary nature of visual speech recognition research fosters collaboration between experts in linguistics, computer science, cognitive psychology, and assistive technology. This interdisciplinary synergy not only advances the field of visual speech recognition but also contributes to the broader landscape of artificial intelligence research and development. By harnessing the capabilities of machine learning and deep learning, visual speech recognition holds the promise of revolutionizing communication, empowering individuals, and fostering a more inclusive and interconnected global society.

Unlocking the Silent Language At its core, lip reading harnesses the visual cues embedded within lip movements to decode spoken language. It taps into the remarkable ability of the human brain to discern phonemes, words, and even complete sentences solely through observing the precise articulations of the lips. This silent language, often imperceptible to the untrained eye, becomes a source of rich information for individuals adept in the art of lip reading.

Lip reading is a vital communication tool for individuals with hearing impairments, enabling them to engage in conversations, educational activities, and various social interactions. It offers a means to bridge the auditory gap, allowing individuals to participate in dialogues that would otherwise be inaccessible. The process of lip reading involves not only the meticulous analysis of lip movements but also the assimilation of facial expressions, contextual cues, and visual context, further enriching the depth of communication.

Applications Across Domains Lip reading's potential impact extends well beyond assisting the hearing-impaired community. Its versatility and applicability render it a valuable asset in various domains, each benefiting from its unique capabilities:

- **Education and Learning:** In educational settings, lip reading complements traditional teaching methods by offering an alternative mode of information transfer. It aids educators in reaching students with diverse learning preferences and caters to those who grasp concepts more effectively through visual stimuli. Additionally, lip reading can facilitate language acquisition and literacy development, particularly for young learners and individuals acquiring new languages.

- **Forensic Analysis:** Lip reading finds applications in forensic investigations, where it aids in deciphering conversations captured by surveillance cameras or audio recordings. The ability to extract meaningful information from visual cues adds a layer of context and comprehension to forensic analyses, potentially influencing legal proceedings.

- **Human-Computer Interaction:** As technology continues to evolve, lip reading holds promise in enhancing human-computer interaction. Gesture-based interfaces and voice-command systems can benefit from the integration of lip reading, enabling more intuitive and context-aware interactions between humans and machines. examples (access control, car driving,..)

- **Medical and Healthcare:** Lip reading's potential extends to the medical field, where it can facilitate effective communication between medical practitioners and patients with speech impairments. Additionally, it may find applications in speech therapy, enabling therapists to fine-tune pronunciation and articulation.

1.2 Challenges and Research Gap

1.2.1 Challenges in Automatic Visual Speech Recognition

The realm of automatic visual speech recognition, though promising, is not devoid of formidable challenges that demand innovative solutions. These challenges, ranging from nuanced subtleties to overarching obstacles, underscore the complexity of deciphering spoken language from visual cues. This section explores the web of challenges that cast a shadow over the path to the widespread adoption of automatic visual speech recognition systems.

Ambiguous Lip Movements One of the central challenges in automatic visual speech recognition lies in the inherent ambiguity of lip movements. While lip reading can be remarkably precise, it is not without its share of details. The visual cues offered by lip articulations often overlap, making it challenging to disentangle phonetic units with absolute certainty. Subtle variations in lip shapes, coupled with the rapid and fluid nature of speech, can lead to misinterpretations and inaccuracies in recognition. Addressing this challenge requires advanced algorithms capable of capturing the subtle differentiators and context cues that distinguish similar lip movements.

Pronunciation Variations Human speech is characterized by a wide spectrum of pronunciation variations, influenced by factors such as accent, dialect, and individual speech patterns. This inherent variability poses a significant hurdle for automatic visual speech recognition systems. The same phoneme can manifest in diverse ways across different speakers, further compounded by variations in articulation speed and emphasis. Adapting recognition models to accommodate these pronunciation variations demands robust training strategies that encompass diverse linguistic contexts and speaker profiles.

Dataset Limitations The effectiveness of automatic visual speech recognition systems hinges on the availability of high-quality and comprehensive datasets. However, the creation of such datasets is fraught with challenges. Factors like dataset size, linguistic diversity, recording conditions, and the representation of real-world scenarios significantly impact the utility of datasets. Existing datasets may be limited in scope, failing to encompass the vast array of speech patterns and visual cues encountered in real-life conversations. The scarcity of large-scale, well-curated datasets that capture the breadth of linguistic and contextual nuances constrains the development and generalization of visual speech recognition models.

Multimodal Fusion Effective visual speech recognition often involves the integration of multiple modalities, such as visual cues, audio signals, and contextual information. The seamless fusion of these modalities to enhance recognition accuracy presents a complex challenge. Achieving synchronized and coherent fusion requires addressing issues related

to temporal alignment, feature integration, and modality weighting. Striking the right balance between modalities while preserving their individual contributions is crucial to harnessing the full potential of multimodal visual speech recognition.

Robustness to Environmental Factors Real-world scenarios introduce a plethora of environmental factors that can compromise the accuracy of visual speech recognition. Variations in lighting conditions, background clutter, occlusions, and camera perspectives introduce noise and distractions that challenge the stability and robustness of recognition models. Developing systems that can operate effectively in diverse and uncontrolled environments, while accounting for these environmental variables, remains an ongoing challenge.

In summary, the challenges in automatic visual speech recognition are a testament to the nature of decoding spoken language from visual cues. Addressing these challenges demands a multidisciplinary approach that combines advancements in machine learning, computer vision, linguistics, and human cognition. As researchers navigate these obstacles, the path to widespread adoption becomes illuminated, offering the promise of enhanced communication and accessibility through automatic visual speech recognition systems.

1.2.2 Need for High-Quality Datasets

The quest for accurate and reliable visual speech recognition models hinges upon the availability of high-quality datasets that encapsulate the rich tapestry of linguistic and visual details. However, the landscape of visual speech recognition is marred by a scarcity of robust and diverse datasets, impeding progress and hindering the development of effective models. This section underscores the indispensable role of datasets in surmounting the challenges inherent to the field.

Dataset Size and Diversity The significance of dataset size cannot be overstated in the realm of visual speech recognition. Larger datasets not only facilitate the training of more robust models but also capture the extensive spectrum of speech patterns, phonetic variations, and contextual cues present in real-world conversations. Diverse datasets, encompassing a wide range of accents, languages, speaking styles, and environmental conditions, enable models to generalize effectively and exhibit enhanced adaptability. The absence of such diversity constrains the model's ability to navigate the intricacies of spontaneous speech and diverse linguistic landscapes.

Realistic Representation A high-quality dataset should strive to emulate the complexities and nuances of authentic communication scenarios. The incorporation of real-world speaking contexts, conversational dynamics, and natural variations in speech articulation is paramount. Such realism enables models to develop an intrinsic understanding of the interplay between visual cues and spoken language, fostering accurate recognition across diverse conversational contexts. The omission of realistic representation limits the model's ability to contend with the multifaceted challenges encountered in practical applications.

Annotation Accuracy The accuracy and precision of dataset annotations play a pivotal role in shaping the effectiveness of visual speech recognition models. Incorrectly annotated data can propagate errors and lead to suboptimal model performance. Accurate labeling of

phonetic units, visemes, and contextual information is imperative to furnish models with ground truth references for training and validation. Moreover, manual or semi-automatic annotation processes must uphold consistency and reliability to instill confidence in the dataset's quality and utility.

Balancing Quality and Quantity Striking a delicate balance between dataset quality and quantity is a challenge that necessitates careful curation. While a large volume of data is desirable, maintaining high-quality standards is equally critical. Curating datasets that encompass a substantial number of utterances while upholding meticulous annotation and representation standards is a demanding task. A dataset that prioritizes both quality and quantity can empower models to transcend limitations and exhibit robust performance across diverse scenarios.

Facilitating Innovation High-quality datasets not only serve as training fodder for existing models but also fuel innovation by enabling the development of novel approaches and architectures. Researchers can leverage rich and diverse datasets to explore new paradigms, devise advanced preprocessing techniques, and experiment with emerging machine-learning techniques. The availability of high-quality datasets acts as a catalyst, fostering a virtuous cycle of research, development, and refinement.

In essence, the dearth of comprehensive and high-quality datasets presents a formidable roadblock toward accurate and effective visual speech recognition. Recognizing the pivotal role of datasets in overcoming the challenges posed by ambiguous lip movements, pronunciation variations, and multimodal fusion underscores the urgency and importance of concerted efforts to curate, enhance, and share datasets that illuminate the path toward transformative advancements in automatic visual speech recognition.

1.3 Research Objectives

The primary objectives of this research are meticulously crafted to address the fundamental challenges in visual speech recognition, especially Arabic VSR, and pave the way for transformative advancements in the field. These objectives serve as the guiding principles that underpin the research endeavor and channel efforts toward achieving tangible and impactful outcomes.

1.3.1 The VSR Framework

This segment focuses on the development and evaluation of a comprehensive Visual Speech Recognition (VSR) framework. The framework encompasses four key services:

1. **Arabic Dataset Creation:** The generation of a robust Arabic dataset serves as the foundational step, providing a diverse and representative collection for subsequent model training and evaluation.

2. **Data Pre-processing & Processing:** This service involves essential pre-processing steps, including mouth detection and tracking, normalization, and data augmentation. The integration of traditional methods like Dlib and modern techniques like YOLO ensures a robust and efficient approach.

3. **Recognition using CNNs and ViT:** The utilization of Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) models forms the core of the recognition

service. The effectiveness and comparative analysis of these models contribute to advancing VSR methodologies.

4. **Framework Integration:** The seamless integration of the developed services into a unified framework allows for a holistic approach to visual speech recognition. This integration aims to balance reliability and efficiency, catering to real-world applications.

1.3.2 Dataset Creation

One of the central thrusts of this research is the creation of a pioneering Arabic audio-visual dataset tailored explicitly for visual speech recognition. This pivotal objective arises from the recognition of a critical void in the availability of suitable datasets, particularly in the context of the Arabic language. The envisaged dataset aspires to be a comprehensive repository that encapsulates the rich diversity of speakers, a varied spectrum of phrases, and an array of real-world conditions. By meticulously curating this dataset, the research endeavors to provide a robust foundation that empowers the development and evaluation of cutting-edge visual speech recognition models. This undertaking resonates with the urgency of overcoming the dataset-related hurdles that have hitherto constrained progress in the field.

Inclusivity and Diversity In pursuit of the broader goal of inclusivity and comprehensive representation, the dataset creation process places a premium on diversity. This objective underscores the imperative to incorporate speakers spanning various age groups, genders, speaking styles, and linguistic backgrounds. By encompassing a wide cross-section of individuals, the research strives to ensure that the dataset mirrors the real-world communication, thus fostering model generalization and adaptability across diverse scenarios.

Phrasal Variety An essential facet of the dataset creation process is the deliberate incorporation of an extensive repertoire of phrases that mirror the linguistic tapestry of everyday conversations. This objective aims to curate a collection of phrases that not only embody the functional richness of spoken language but also exhibit the nuances and challenges inherent in lip reading. By encompassing a diverse array of phrases, from greetings to statements to questions, the dataset aspires to be a versatile resource that underpins multifaceted research endeavors and practical applications.

Robustness to Conditions Recognizing the significance of environmental conditions in shaping the efficacy of visual speech recognition models, this research endeavors to infuse the dataset with resilience to varying conditions. The objective of robustness entails capturing videos under different lighting conditions, indoor environments, and potential acoustic variations. By diversifying the conditions under which data is captured, the dataset aims to bolster model adaptability and equip researchers with a toolset to address real-world challenges.

In essence, the primary objectives of this research crystallize into a comprehensive roadmap aimed at surmounting key challenges in visual speech recognition. From the creation of a novel Arabic audio-visual dataset to the infusion of inclusivity, diversity, phrasal variety, and robustness, these objectives form the bedrock upon which the research endeavors to advance the frontiers of automatic visual speech recognition.

1.3.3 Data Pre-processing & Processing:

This pivotal service encompasses several crucial steps in preparing and refining the data for effective visual speech recognition.

Mouth Detection and Tracking: The first subcomponent involves leveraging a dual-method approach for accurate mouth region identification and tracking. Traditional methods like Dlib provide reliability, while modern real-time techniques like YOLO contribute efficiency to the process.

Normalization: Another integral step is standardizing the data to a consistent format, ensuring uniformity, and facilitating optimal model training. Normalization techniques are applied to enhance the reliability of the input data, providing a standardized foundation for subsequent stages.

Data Augmentation: To diversify and expand the dataset, the third subcomponent introduces data augmentation techniques such as vertical and horizontal shifts, and random zoom. This step aims to enhance model generalization and robustness by introducing variability into the training data.

Integration of Pre-processed Data: Finally, the pre-processed data, enriched with accurately detected and tracked mouth regions, undergoes seamless integration into the subsequent stages of the VSR framework. This integration ensures that the refined data is effectively utilized in model training and further analysis, forming a cohesive and streamlined data processing pipeline.

1.3.4 Development of Visual Speech Recognition Models

At the core of this research lies the ambitious endeavor to propel the advancement of automatic visual speech recognition through the development and rigorous evaluation of sophisticated deep learning models. With the newly created Arabic audio-visual dataset as the crucible of exploration, this objective aspires to unearth novel insights, techniques, and architectures that transcend the boundaries of traditional methodologies.

Exploration of Two Distinct Approaches Central to the pursuit of this objective is the exploration and delineation of two distinct and potent avenues: convolutional neural networks (CNN) and vision transformers. The research seeks to harness the transformative power of these paradigms, leveraging their respective strengths and capabilities in unravelling the detail dynamics of visual speech recognition.

Convolutional Neural Networks (CNN) The first prong of this exploration involves the systematic construction and training of convolutional neural network models tailored explicitly for visual speech recognition. Harnessing the innate ability of CNNs to extract spatial hierarchies, patterns, and features from visual data, this approach aspires to endow models with the discernment to decipher the subtle details of lip movements. The research explores nuanced architecture design, hyperparameter tuning, and training protocols that culminate in the realization of a potent CNN-based visual speech recognition model.

Vision Transformers The second facet of exploration embarks on the terrain of vision transformers, a paradigm that has emerged as a formidable force in image understanding and analysis. Embracing the inherent capacity of transformers to capture long-range dependencies and contextual information, this approach aims to foster a new paradigm in visual speech recognition. Through the careful orchestration of attention mechanisms, positional encodings, and self-attention architectures, the research endeavors to craft a visionary vision transformer-based model that redefines the boundaries of accuracy and efficiency in visual speech recognition.

Quantitative and Qualitative Evaluation Crucially, this objective transcends mere model development, venturing into the realm of comprehensive evaluation. The proposed models will be subjected to rigorous quantitative assessments, delving into performance metrics such as accuracy, precision, recall, and F1-score. Equally vital is the qualitative evaluation, where the models' ability to encapsulate linguistic nuances, adapt to diverse speakers, and navigate real-world conditions will be scrutinized under the lens of practicality and efficacy.

In summary, the development of visual speech recognition models stands as a linchpin objective that weaves together innovation, exploration, and evaluation. Through the symphony of convolutional neural networks and vision transformers, this research aspires to chart a transformative trajectory in automatic visual speech recognition, culminating in models that can decipher the unspoken language of lips with unparalleled acumen.

1.3.5 Modern Standard Arabic (MSA)

The research context is carefully bound to concentrate on Modern Standard Arabic (MSA), one of the prominent linguistic variants within the Arabic language spectrum. While Arabic unfolds as a tapestry of diverse dialects and accents, this study consciously narrows its lens to MSA. This strategic selection is underpinned by a nuanced rationale that underscores both the advantages and limitations of this approach.

The research's strategic choice to centre on Modern Standard Arabic (MSA) reflects a balance between fostering linguistic unity and recognizing the multitudinous facets of Arabic linguistic heritage.

1.4 Thesis Organization

The remainder of this thesis is structured as follows:

Chapter 2: Theoretical Background This chapter provides a Theoretical Background by elucidating foundational concepts, frameworks, and relevant literature essential for contextualizing the research problem. It aims to establish a solid theoretical foundation, laying the groundwork for the subsequent chapters.

Chapter 3: Lip Reading and Visual Speech Recognition This chapter provides an in-depth exploration of lip reading, visual speech recognition approaches, and the significance of datasets in this context.

Chapter 4: Dataset Creation Detailed information about the creation process of the proposed Arabic audio-visual dataset (BlidAVS10) is presented in this chapter.

Chapter 5: Visual Speech Recognition Models The development and evaluation of deep learning models, including convolutional neural networks and vision transformers, are discussed in this chapter.

Chapter 6: Experimental Results and Discussion The experimental results of the visual speech recognition models are analyzed, discussed, and compared in this chapter.

Chapter 7: Conclusion and Future Work The final chapter summarizes the research findings, emphasizes key contributions, and outlines potential avenues for future research.

1.5 Publications

[7] Baaloul, A., Benblidia, N., Reguieg, F. Z., Bouakkaz, M., Felouat, H. (2024). An arabic visual speech recognition framework with CNN and vision transformers for lipreading. Multimedia Tools and Applications, 1-35.

Part II

State of the Art

Chapter 2

Theoretical Background

2.1 Introduction

In this section, we establish the foundational framework for our research on automated lip-reading. We provide an in-depth exploration of fundamental concepts and techniques, delving into key areas such as computer vision, image processing, deep learning, and object detection and tracking algorithms. This groundwork is essential for a comprehensive understanding of the subsequent discussions and analyses in our study.

2.2 Computer Vision and Image Processing

Computer vision (CV) is an interdisciplinary field that plays a pivotal role in enabling computers to interpret and understand visual information from the surrounding environment. At its core, computer vision aims to replicate the remarkable human ability to perceive, analyze, and comprehend visual data, making it an indispensable technology with a wide range of applications. One of the fundamental components of computer vision is image processing, a set of techniques that involve manipulating and analyzing digital images to extract valuable information, enhance visual quality, and facilitate various tasks.

2.2.1 Image Processing Techniques

Image processing is a crucial subset of computer vision that involves a series of operations performed on digital images to improve their quality, enhance specific features, or extract meaningful information. These techniques serve as building blocks for more complex computer vision applications, such as lip reading systems.

1. **Image Enhancement:** Enhancement techniques aim to improve the quality and clarity of images, making them more suitable for subsequent analysis. Common methods include contrast stretching, histogram equalization, and noise reduction.

2. **Image Filtering:** Image filtering involves applying convolutional filters to images, resulting in effects such as blurring, sharpening, and edge detection. These filters highlight specific features and help in preprocessing image data.

3. **Feature Extraction:** Feature extraction techniques identify and isolate specific characteristics or patterns within an image. In the context of lip reading, these methods can be used to isolate the mouth region and identify relevant visual cues.

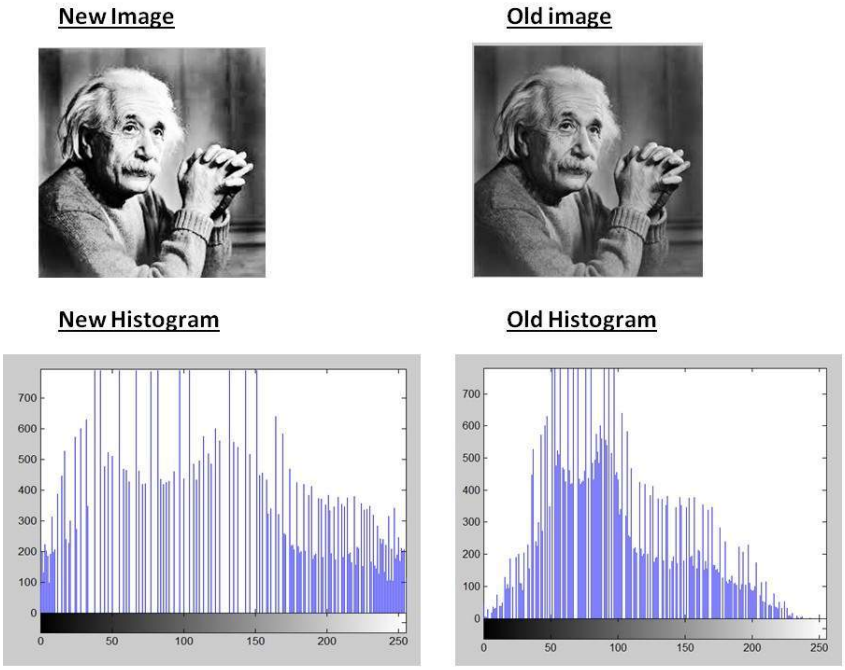


FIGURE 2.1: histogram equalization.

4. **Image Segmentation:** Image segmentation divides an image into distinct regions based on common properties, enabling the separation of objects from the background. This is particularly useful for isolating the lips and distinguishing them from the rest of the face.

2.2.2 Role in Lip Reading Systems

In the domain of visual speech recognition and lip reading, computer vision and image processing techniques play a pivotal role in several key aspects:

- **Mouth Region Extraction:** Accurate lip reading systems require the precise extraction of the mouth region from visual data. Computer vision methods enable the identification of the mouth's boundaries, even in challenging conditions, by leveraging techniques such as edge detection and region segmentation.
- **Feature Representation:** Image processing aids in the extraction of relevant visual features from the mouth region. These features capture critical information about lip movements, shapes, and articulations that are essential for accurate speech recognition.
- **Preprocessing and Normalization:** Computer vision techniques help in preprocessing raw image data by reducing noise, adjusting lighting conditions, and enhancing visual quality. This normalization step ensures that the input data is standardized and optimal for subsequent analysis.
- **Data Augmentation:** To train robust and generalized lip reading models, data augmentation techniques are employed. Computer vision methods can introduce variations in lighting, pose, and other factors to artificially expand the training dataset.

In summary, computer vision and image processing provide the foundational tools and methodologies for processing visual information, a cornerstone of successful lipreading systems. These techniques enable the extraction of meaningful features, enhance data quality, and contribute to accurate and reliable visual speech recognition.

The next section will explore the deep learning paradigms, including Convolutional Neural Networks (CNNs) and ViTs, which leverage these processed visual features for advanced analysis and recognition tasks.

2.3 Role of SVM and HMM

Support Vector Machines (SVM) and Hidden Markov Models (HMM) are two traditional classification methods that have been applied to visual speech recognition tasks. These methods play distinct roles in processing the features extracted from lip images and addressing the temporal nature of speech.

Support Vector Machines (SVM) Support Vector Machines (SVM) are a class of supervised learning algorithms that excel in binary classification tasks. In the context of visual speech recognition, SVMs are used to assign phonemes or visemes to the extracted handcrafted features. SVMs aim to find a hyperplane that best separates the feature space into distinct classes, optimizing the margin between data points of different categories. The decision boundary learned by SVMs facilitates the classification of lip images into their corresponding phonetic units.

However, the effectiveness of SVMs in handling sequential data and capturing temporal dependencies is limited. Lip reading involves analyzing the dynamic changes in lip movements over time, which are better captured by models capable of handling sequential patterns. This limitation led to the exploration of Hidden Markov Models (HMMs) to address the temporal aspect of lip sequences.

Hidden Markov Models (HMM) Hidden Markov Models (HMMs) are probabilistic models widely used for sequential data analysis. HMMs are particularly suitable for tasks that involve temporal patterns, such as speech recognition. In the context of lip reading, HMMs are employed to model the transitions between different phonetic states over time. Each state in the HMM corresponds to a specific phoneme or viseme, and the transitions between states capture the temporal relationships between these units.

HMMs use the observed sequences of handcrafted features extracted from lip images to estimate the most likely sequence of phonetic units. The model incorporates both the emission probabilities of features from each state and the transition probabilities between states. Training HMMs involves using methods like the Expectation-Maximization (EM) algorithm to learn the model parameters that maximize the likelihood of observed data.

While SVMs and HMMs have been instrumental in the early stages of visual speech recognition research, the advent of deep learning has led to the development of more powerful models capable of learning hierarchical features and capturing complex temporal dependencies. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have revolutionized the field, enabling end-to-end training and outperforming traditional methods in various aspects.

2.4 Overview of Deep Learning

Deep learning is a subfield of machine learning that focuses on neural networks with multiple layers, known as deep neural networks. These networks leverage hierarchical representations to learn complex patterns and features from data. The depth of the network allows it to automatically extract hierarchical features, enabling it to capture intricate relationships within the input data, Unlike traditional machine learning models that rely on manually engineered features, making them highly effective in tasks involving complex patterns and large datasets.

Central to deep learning is the concept of artificial neural networks, inspired by the structure and functioning of the human brain. These networks consist of interconnected nodes organized into layers, including an input layer, one or more hidden layers, and an output layer. The connections between nodes, represented by weights, are adjusted during training to optimize the network's performance.

Deep learning has demonstrated remarkable success in various domains, including computer vision, natural language processing, and speech recognition. Convolutional Neural Networks (CNNs) excel in image-related tasks, Recurrent Neural Networks (RNNs) are effective for sequential data, and Transformer architectures have revolutionized natural language processing.

Key components of deep learning include activation functions, loss functions, and optimization algorithms. Activation functions introduce non-linearity, enabling the model to learn complex mappings. Loss functions measure the difference between predicted and actual values, guiding the training process. Optimization algorithms adjust model parameters to minimize the loss function.

Understanding these fundamental aspects of deep learning is crucial for appreciating its applications and advancements in various fields.

2.4.1 Convolutional Neural Networks (CNNs)

CNN (Convolutional Neural Network) is a class of deep learning algorithms pioneered by LeCun [47], extensively utilized in computer vision tasks like image classification, object detection, and facial recognition.

CNNs are renowned for their capability to achieve cutting-edge performance in image recognition tasks and their adeptness in handling high-dimensional input data. They excel in recognizing objects in images, identifying faces, detecting anomalies in medical images like cancer, and powering autonomous vehicle navigation. Moreover, CNNs find applications in natural language processing, speech recognition, and other domains involving sequential data analysis.

The core concept underlying CNNs is their ability to automatically extract pertinent features from raw data without necessitating manual feature engineering. This attribute renders them highly adaptable and suitable for diverse tasks, even in scenarios with limited labeled data. Feature extraction from input data, such as images, is accomplished by convolving small matrices known as filters over the input, performing mathematical operations at each position to generate output values. Subsequent network layers then process these resulting feature maps to progressively construct a representation of the input.

- **Convolutional layer**

The convolutional layer is pivotal in convolutional neural networks (CNNs), applying filters (kernels) to the input image or feature map for feature extraction across various spatial locations. By sliding a small filter over the input and computing the dot product with the local receptive field at each location (Figure 2.2), it generates a feature map indicating the presence or absence of specific visual patterns. The stacking of multiple convolutional layers allows the model to learn diverse features, enabling the extraction of increasingly complex patterns from the input data. Convolutional layers excel in the automatic extraction of hierarchical representations, making them highly effective for image-related tasks like recognition and detection, emphasizing relevant features for accurate predictions. Their parameter-sharing efficiency minimizes the overall model parameters, promoting generalization to new, unseen data.

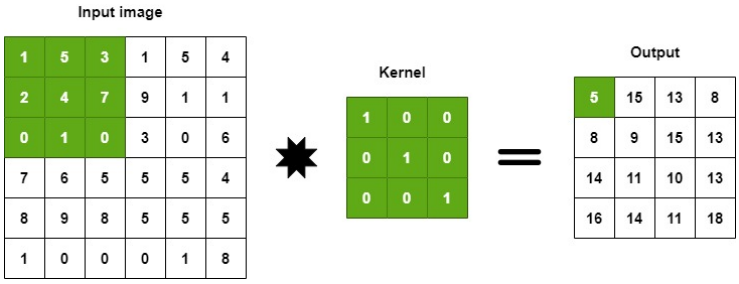


FIGURE 2.2: The convolution operation

Example of *the convolution* operation that results from an output matrix by sliding a filter 3x3 over the pixels of the input image 6x6, performing a multiplication and addition operation at each location, and storing the result in the output matrix. The resulting image may be filtered, blurred, or sharpened, depending on the characteristics of the kernel used.

- **MaxPooling** MaxPooling, a commonly utilized pooling layer in convolutional neural networks (CNNs) [9], serves to decrease the spatial dimensions of the input image, enhancing the network's learning efficiency and mitigating overfitting. In this layer, a fixed-size window (typically 2x2) slides over the input image, selecting the maximum value within each window as the output (Figure 2.3). The stride, equal to the window size, ensures no overlap between adjacent windows.

  MaxPooling operates by dividing the input into rectangular or square regions, outputting the maximum value for each region based on the pool size and stride. This layer achieves translation invariance, meaning the output remains unaffected by the object's location in the input image. Additionally, it achieves spatial invariance, as the output remains consistent even if the object varies slightly in size. Crucially, MaxPooling reduces feature map dimensionality, curbing overfitting by preventing the network from memorizing training data.

- **Dense layer** The dense layer, or fully connected layer, connects every neuron in the current layer to the subsequent one, processing input values through multiplication by a weight matrix, the addition of a bias term, and an activation function. Its capability to learn complex mappings suits tasks like image classification, natural language processing, and time-series analysis, particularly for high-dimensional and nonlinear data. The number of neurons in the dense layer governs the complexity of the learned function, influencing the network's capacity to discern patterns and the risk of overfitting.

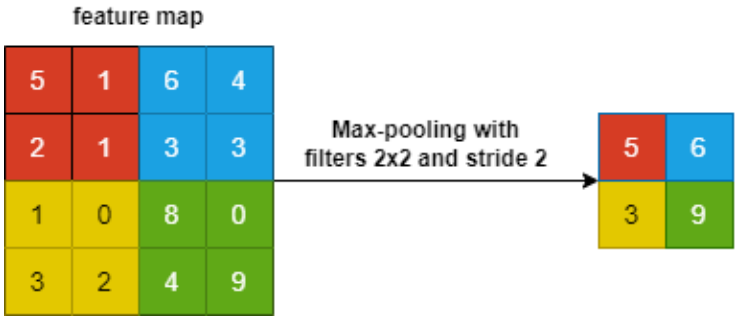


FIGURE 2.3: The max-pooling technique.

In this example, we show the max-pooling technique used to help reduce the dimensionality of the feature map while retaining the most important information by using a window size of 2x2 and stride of 2, this means we take the maximum value within each 2x2 window and discard the other values

- **Dropout layer** The dropout layer, a widely-used regularization technique in deep learning, combats overfitting by randomly setting a percentage of neurons in a layer to zero during each training iteration [94]. This stochastic regularization forces the network to learn redundant data representations, enhancing robustness and preventing individual neurons from exerting disproportionate influence. Typically applied after a dense or fully connected layer and followed by another dense layer, the Dropout layer is deactivated during inference, ensuring all neurons contribute to predictions. While effective for preventing overfitting, prudent use is crucial to avoid underfitting and maintain optimal model performance.

- **Batch normalization layer** Batch normalization is a technique used in deep learning to improve the training of neural networks. It is typically applied to the output of a convolutional or fully connected layer. It helps to normalize the input to the next layer by normalizing the input of each layer by subtracting the mean and dividing by the standard deviation of the inputs [41].

  Batch normalization has several advantages, including reducing the network's sensitivity to initializing its weights and biases, making it less likely to get stuck in local minima during training. It can also mitigate the effects of the vanishing and exploding gradient problems that can occur during back-propagation, faster convergence during training, and improved generalization performance.

CNNs in Lip Reading: CNNs have demonstrated remarkable prowess in extracting detailed visual patterns from lip images, making them indispensable in automated lip-reading systems. By leveraging the hierarchical features learned through convolutional and pooling layers, CNNs can discern subtle lip movements and articulations that encode spoken language. The ability of CNNs to automatically learn relevant features from raw visual data contributes to the accuracy and efficacy of automated lip reading, advancing the field of visual speech recognition.

The subsequent section will introduce an alternative approach in the realm of deep learning—ViTs—and discuss its applicability and advantages in the context of automated lip reading.

2.4.2 Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs) [52] represent a class of neural networks specifically tailored for processing sequential data, leveraging a memory-like structure. Diverging from traditional feedforward neural networks, RNNs feature cyclic connections, enabling them to retain information from previous inputs in an internal state (Figure 2.4 below shows the RNN's model).

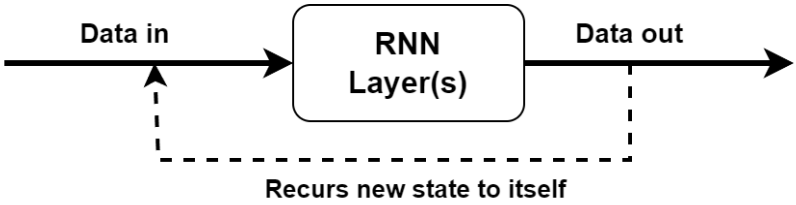


FIGURE 2.4: Recurrent Neural Network Model.

The hallmark of RNNs lies in their capability to capture dependencies and interrelations within sequential data, making them particularly effective for tasks such as natural language processing, speech recognition, and time-series analysis. RNNs sequentially process input elements while maintaining a hidden state that encapsulates information from preceding inputs.

Nevertheless, standard RNNs grapple with learning long-range dependencies, encountering challenges like vanishing or exploding gradients. To overcome these hurdles, evolved architectures like Long Short-Term Memory (LSTM) [39] and Gated Recurrent Unit (GRU) have been introduced. These variants integrate mechanisms for selective information updating and forgetting, augmenting their proficiency in capturing long-term dependencies.

2.4.3 Long Short-Term Memory (LSTM)

To address challenges like vanishing and exploding gradients, Hochreiter and Schmidhuber [39] introduced Long Short-Term Memory (LSTM). A key advantage of LSTM is its cell state, which facilitates the storage and controlled flow of information from input to output. Figure 2.5 illustrates the fundamental architecture of an LSTM cell, comprising essential components such as the input gate, output gate, forget gate, and update gate. The forget gate determines which information to discard from past memory units, the input gate regulates which information to incorporate into the neuron, the output gate produces new long-term memory, and the update gate modifies the cell. These components collaborate in a systematic manner, processing input sequences, managing short-term and long-term memories, and generating output sequences at each time-step.

The following equation is a mathematical expression of the input gate which determines what information must be transferred to the cell:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

Whereas the following equation is a mathematical expression of the forget gate which determines which information to be neglected:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

The update gate updates the cell state which is mathematically expressed by the following equations:

$$c_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$$

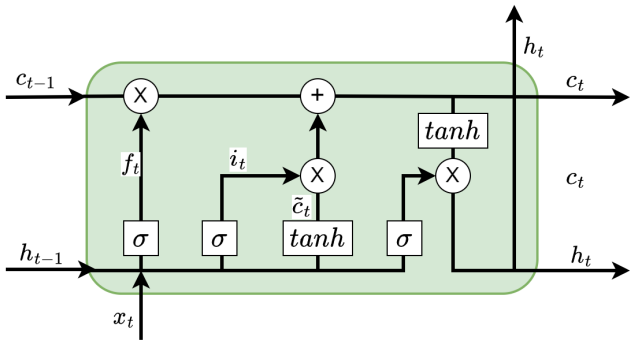


FIGURE 2.5: Schematic representation of Long-Short Term Memory (LSTM) cell.

$$c_t = f_t c_{t-1} + i_t c_t$$

The output gate is responsible for updating the output and is given by the following equation; the output gate is also responsible for updating the hidden layer of the previous time-step:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

$$h_t = ot \tanh(c_t)$$

2.4.4 Gated Recurrent Unit (GRU)

Chung et al. [15] introduced Gated Recurrent Units (GRUs) as a simplified version of the LSTM cell, offering reduced training time coupled with enhanced network performance (Figure 2.6). Functionally akin to LSTMs, GRUs merge the forget and input gates into a single update gate, streamlining operations. Additionally, GRUs consolidate hidden and cell states into one, effectively halving the total number of gates compared to LSTM. This architectural simplicity has popularized GRUs as a concise alternative to LSTM cells. The two primary gates in a GRU are the update gate and the reset gate. The hidden state of the GRU is represented by the following equation:

$$h_t = (1 - z_t) * h_{t-1} + z_t * h_t$$

The following equation represents the update gate and determines how much of the GRU unit gets updated:

$$z_t = \sigma(W_z * [h_{t-1}, x_t])$$

The reset gate is given by the following equation:

$$r_t = \sigma(W_r * [h_{t-1}, x_t])$$

The hyperbolic tangent function of the reset gate is called a new remember gate, which is described by the following function:

$$h_t = \tanh(W * [r_t * h_{t-1}, x_t])$$

In pivotal applications like language modeling, machine translation, and speech recognition, RNNs have played foundational roles. Despite the emergence of newer architectures like Transformers in the dynamic landscape of deep learning, RNNs remain indispensable for tasks demanding nuanced sequential context.

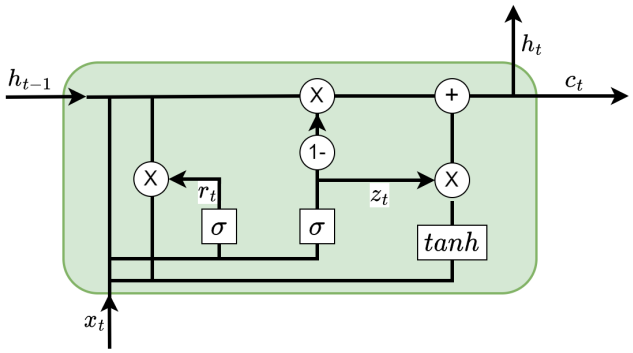


FIGURE 2.6: Schematic representation of Gated Recurrent Unit (GRU) cell.

2.4.5 Transformers and Vision Transformers

Transformers have redefined the landscape of various domains, particularly in natural language processing, due to their exceptional ability to capture complex relationships in data. In this section, we explore the transformative architecture of Transformers [103] and its adaptation to visual tasks, exemplified by ViTs. While initially designed for text-based applications, the extension of Transformers to visual inputs holds immense promise in enhancing lip-reading systems by leveraging their capability to capture long-range dependencies and contextual information within sequences.

2.4.5.1 The Transformer Architecture

- **Self-Attention Mechanism:** At the core of the Transformer is the self-attention mechanism, allowing the model to weigh the importance of different elements in a sequence dynamically. This holistic processing of input data is particularly advantageous for capturing contextual information and dependencies in sequences.

- **Multi-Head Self-Attention:** To enhance the expressive power of self-attention, Transformers employ a multi-head mechanism. This involves running the self-attention mechanism in parallel multiple times, each with different learned linear projections. Multi-head self-attention enables the model to focus on different aspects of the input sequence concurrently, allowing for more comprehensive feature extraction.

- **Feed-Forward Neural Networks:** In addition to self-attention, Transformers integrate feed-forward neural networks as a crucial component. This layer follows the self-attention mechanism and consists of fully connected layers, providing the model with the capacity to capture complex patterns and relationships within the learned representations.

  For a visual representation, refer to Figure 2.7.

- **Applications:** The Transformer architecture has demonstrated exceptional performance across various natural language processing tasks, including but not limited to:

  - **Machine Translation:** Transformers excel in capturing contextual information for accurate translation between languages.

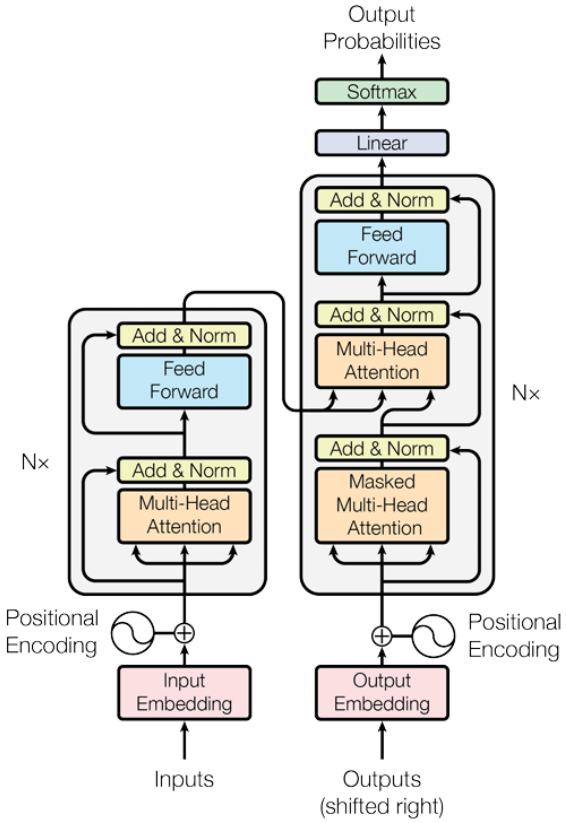


FIGURE 2.7: The Transformer - model architecture.

  - **Text Generation:** The ability to model long-range dependencies makes Transformers effective for generating coherent and contextually relevant text.
  - **Language Understanding:** Transformers have been successfully employed in tasks such as sentiment analysis, named entity recognition, and language understanding, showcasing their versatility.

- **Transformer Variants:** Since its introduction, the Transformer architecture has evolved, leading to various variants tailored for specific tasks and data modalities. Notable variants include BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), and (Text-to-Text Transfer Transformer), each contributing to advancements in natural language processing and beyond.
- **Conclusion:** The Transformer architecture has revolutionized the field of deep learning by introducing a scalable and efficient mechanism for capturing dependencies in sequential data. Its self-attention mechanism and multi-head approach provide a powerful framework that has been adapted and extended in numerous subsequent models. As research continues, Transformers remain at the forefront of cutting-edge developments in natural language processing and beyond.

2.4.5.2 Vision Transformer

The Vision Transformer (ViT) is an innovative model architecture designed for image recognition tasks, leveraging the Transformer architecture initially developed for natural

language processing (NLP). ViT transforms 2D images into sequences of flattened patches and processes them using a standard Transformer encoder employed in NLP tasks.

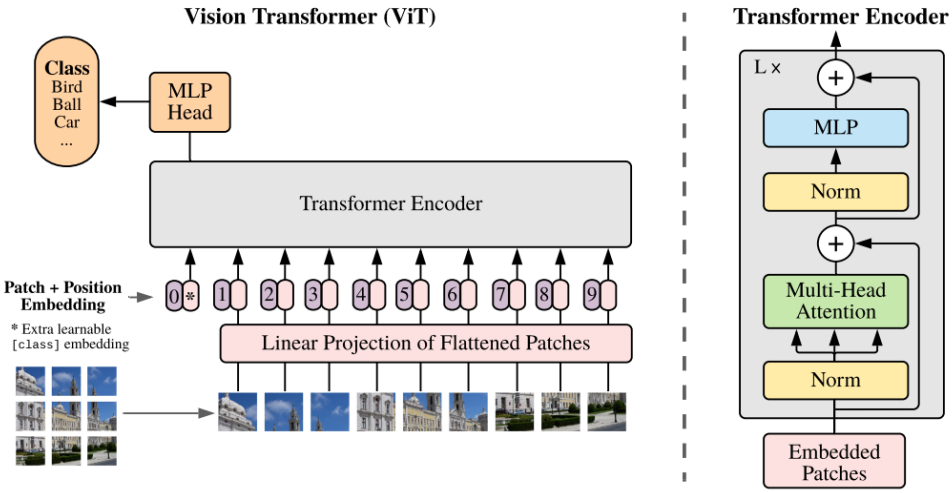


FIGURE 2.8: Overview of the ViT model.

The image is divided into fixed-size patches, each linearly embedded and augmented with position embeddings. The resulting sequence of vectors is then fed into a standard Transformer encoder. For classification, an additional learnable "classification token" is appended to the sequence. The illustration of the Transformer encoder draws inspiration from [103].

To handle 2D images, See Figure 2.8 ViT divides the image into non-overlapping patches of a fixed size, flattening each patch into a 1D vector. These patch embeddings then undergo processing in the Transformer encoder, comprising multiple layers of self-attention and feedforward neural networks. The self-attention mechanism enables the model to focus on different parts of the image while processing each patch, while the feedforward networks capture non-linear relationships between patches.

ViT employs a pre-training task known as "masked image modeling" to learn meaningful representations of image patches. In this task, a random subset of patches is masked, and the model is trained to predict the masked patches based on the remaining patches. This pre-training aids the model in learning valuable features transferable to downstream image recognition tasks.

In summary, ViT presents a straightforward yet scalable approach to image recognition, demonstrating remarkable performance when coupled with pre-training on extensive datasets. ViT has achieved state-of-the-art results in various image recognition benchmarks while demanding fewer computational resources for training compared to traditional convolutional neural networks.

The subsequent section delves into the practical implementation of ViTs within the context of our proposed visual speech recognition models, shedding light on their effectiveness and contributions to accurate lip-reading systems.

2.4.6 Object Detection Algorithms

The field of object detection, particularly within the realm of Neural Network models, stands as a significant sub-discipline in deep learning. Its primary objective is to discern the presence of single or multiple objects within images or video streams. The applications

of object detection span various domains, encompassing object tracking, video surveillance, pedestrian detection, anomaly detection, people counting, self-driving cars, face detection, and beyond. Over the years, numerous machine learning and deep learning techniques have been proposed, each with its distinct advantages and limitations.

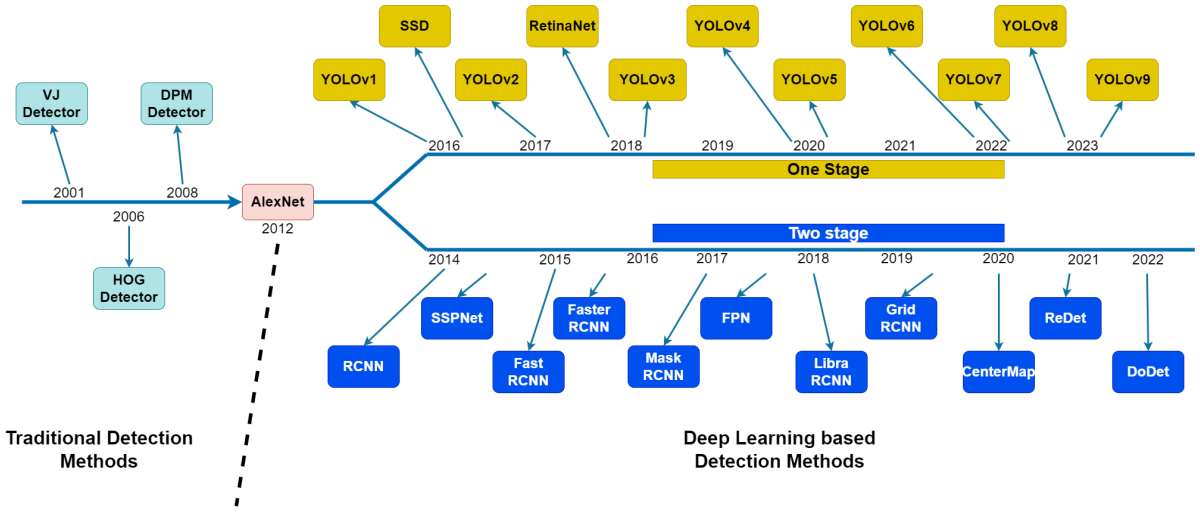


FIGURE 2.9: Milestones of object detection

Various methodologies have been developed to tackle the challenges posed by object detection. Haar Cascades [104], utilizing Haar-like features and a cascade of classifiers, offers a fast approach, although it may have limitations in accuracy, especially for complex scenes. Histogram of Oriented Gradients (HOG) [22] computes gradients and histograms of image gradients, proving effective in capturing object shapes and contours, often paired with support vector machines. Region-based CNNs (R-CNN) [34] introduced the concept of region proposals, while Fast R-CNN [33] integrated region proposal generation into the network for improved speed. Faster R-CNN [85] took a step further by introducing a Region Proposal Network (RPN) within the network, achieving end-to-end object detection with enhanced speed and accuracy. You Only Look Once (YOLO) [83, 81, 82] divided the image into a grid, predicting bounding boxes and class probabilities for each grid cell, offering real-time object detection in a single pass through the network. Single Shot Multibox Detector (SSD) [53] balanced speed and accuracy by utilizing default bounding boxes at different aspect ratios and scales. RetinaNet addressed class imbalance through the introduction of the Focal Loss, ensuring high accuracy by focusing on challenging examples. EfficientDet efficiently scaled networks for object detection, achieving competitive performance with smaller and faster models. CenterNet, with its focus on keypoint estimation, accurately predicted object centers and regressed bounding box dimensions. These algorithms represent a continuum of advancements, each catering to specific needs in terms of speed, accuracy, and efficiency, allowing for a tailored approach depending on the requirements of the application at hand, See Figure 2.9.

These models represent cutting-edge solutions in the pursuit of efficient and accurate object detection, contributing to advancements in diverse applications.

2.4.6.1 Traditional Methods: Haar and HOG Approaches

Traditional object detection relies on established methods such as the Haar Cascade Classifiers and the Histogram of Oriented Gradients (HOG). These time-tested approaches

have played a crucial role in computer vision, offering effective means of detecting objects within images.

- **Haar Cascade Classifiers**

  The Haar method employs Haar-like features, utilizing rectangular filters to analyze intensity variations in different regions of an image. Recognizing specific patterns associated with the object of interest, Haar Cascade Classifiers create a cascade of classifiers for efficient detection.

FIGURE 2.10: AdaBoost feature analysis: contrasting eye and cheek brightness and eye-nose intensity.

AdaBoost selected the first and second features for analysis. These two features are displayed in the top row and then superimposed on a standard training face in the bottom row. The first feature calculates the variation in brightness between the area around the eyes and a region spanning the upper cheeks. This feature takes advantage of the fact that the eye region tends to be darker than the cheeks. On the other hand, the second feature compares the brightness levels in the eye region with the intensity across the bridge of the nose.

- **Histogram of Oriented Gradients (HOG)**

  HOG represents the distribution of gradient orientations in an image, dividing it into small cells to capture local object shape and structure. This traditional method excels at detecting edges and contours.

Acknowledging the traditional nature of Haar and HOG methods underscores their historical significance in the landscape of object detection. These approaches laid the groundwork for subsequent advancements in the field.

2.4.6.2 Fast R-CNN

Fast R-CNN is a significant advancement in object detection, addressing limitations observed in prior approaches like R-CNN and SPPnet. This method excels in achieving heightened detection accuracy while substantially improving training and testing speeds. Key innovations of Fast R-CNN include the incorporation of a Region of Interest (RoI) pooling layer. This layer facilitates efficient back-propagation and enables the direct evaluation of object proposal mean Average Precision (mAP). Notably, Fast R-CNN streamlines the training process by replacing the multi-stage pipeline of R-CNN with a more efficient single-stage training approach, optimizing both time and space utilization.

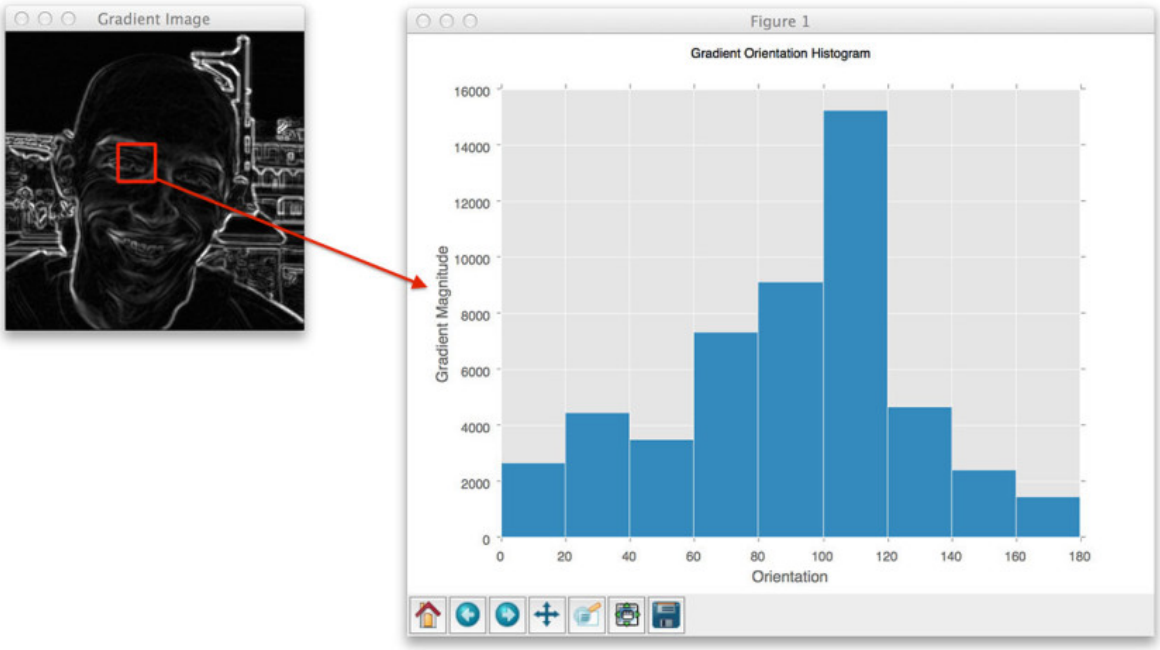


FIGURE 2.11: An instance of calculating a histogram of oriented gradients for a specific image region.

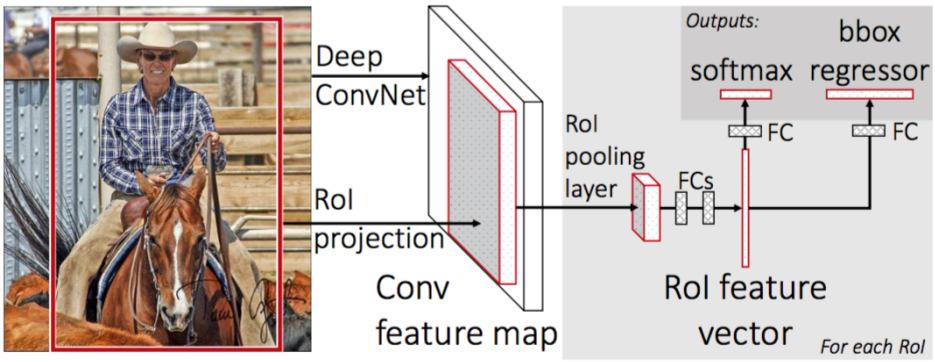


FIGURE 2.12: Architecture of Fast R-CNN, showcasing the RoI pooling layer, sibling layers for bounding-box regression and category classification, and the streamlined single-stage training process.

Moreover, Fast R-CNN introduces a sibling layer for bounding-box regression and category classification. It further refines model performance by fine-tuning all network weights through back-propagation. These architectural enhancements collectively contribute to the efficiency and effectiveness of the Fast R-CNN method.

For a more comprehensive understanding, refer to the visual representation of the Fast R-CNN architecture provided in Figure 2.12. This image illustrates the key components of the network, including the Region of Interest pooling layer, bounding-box regression, and category classification, enhancing the clarity of the method's structural innovations.

2.4.6.3 Faster R-CNN

The Faster R-CNN method introduces a Region Proposal Network (RPN) that efficiently shares convolutional layers with state-of-the-art object detection networks. This innovative approach enables the computation of proposals at test-time with a minimal cost, as low as 10ms per image. The RPN is trained end-to-end through back-propagation and stochastic gradient descent, employing an "image-centric" sampling strategy for network optimization.

To accommodate varying object sizes, the RPN utilizes a set of bounding-box regressors, with each regressor tailored to a specific scale and aspect ratio. This design allows the prediction of boxes of different sizes, even though the features used for regression remain of a fixed size/scale. The RPN's elegant solution ensures nearly cost-free proposal computation, particularly when integrated with the detection network.

A notable strength of the RPN lies in its graceful behavior as the number of proposals decreases, contributing to its high ultimate detection mean Average Precision (mAP) even with as few as 300 proposals. This robustness can be attributed to the classification term of the RPN. The two-stage cascade of the RPN, involving class-agnostic proposals and class-specific detections, sets it apart from the one-stage, class-specific detection pipeline proposed in the OverFeat paper.

In summary, the Faster R-CNN method introduces a novel Region Proposal Network, significantly enhancing the efficiency and accuracy of object detection systems. This advancement marks a noteworthy contribution to the intersection of computer vision and deep learning (**Figure 2.13**).

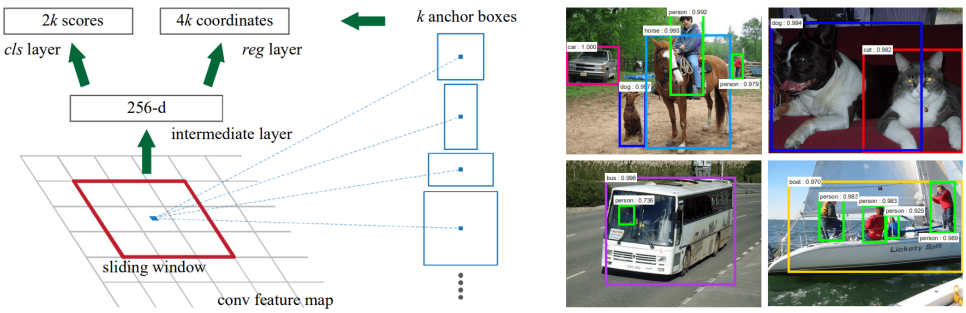


FIGURE 2.13: Architecture of the Faster R-CNN method.

2.4.6.4 You Only Look Once (YOLO)

YOLO, or "You Only Look Once," is a real-time object detection system developed by Joseph Redmon et al. in 2016. It introduced an end-to-end approach for object detection, accomplishing the task with a single pass, a departure from previous methods using sliding windows or two-step processes, See (Figure 2.14).

The distinctive feature of YOLO is its capability to detect objects in real time with high accuracy by dividing the input image into a grid. It predicts bounding boxes and class probabilities for each grid cell, enabling the detection of multiple objects in a single pass. This approach is renowned for its speed and efficiency in real-time applications.

Throughout its evolution, YOLO has undergone various versions, including YOLOv1, YOLOv2, YOLOv3, YOLOv4, YOLOv5, YOLOv6, YOLOv7,YOLOv8, and YOLOv9. Variants such as YOLO-NAS and YOLO with Transformers have expanded their capabilities,

contributing to their widespread use in robotics, driverless cars, video monitoring, action recognition, sports analysis, and human-computer interaction.

Continuous refinement has positioned YOLO as a central real-time object detection system, adapted for diverse applications. Ongoing research focuses on enhancing its performance for specific use cases, such as small object detection in autonomous vehicles, intelligent surveillance, and multi-task learning for various detection and classification tasks.

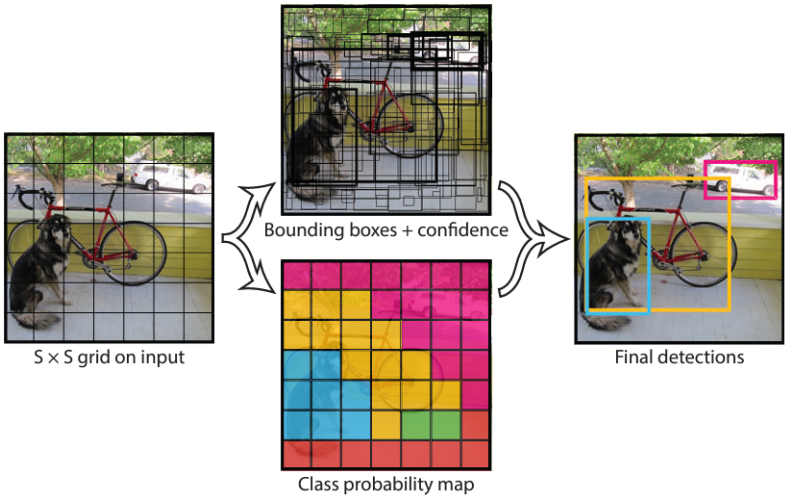


FIGURE 2.14: The Yolo model treats detection as a regression problem, predicting bounding boxes, confidence scores, and class probabilities for each grid cell in an S×S grid. Predictions are encoded into an S×S×(Bx5+C) tensor, with B representing bounding boxes and C representing class probabilities.

Evolution of YOLO Versions: The YOLO (You Only Look Once) object detection framework has undergone significant evolution with each version, introducing improvements in accuracy, speed, and adaptability. A timeline of key YOLO versions is depicted in (**Figure 2.15**). Here is a concise overview of these versions:

- YOLOv1
  - Introduced real-time object detection with a single pass through the network.
  - Implemented grid division, predicting bounding boxes and class probabilities for each cell.
  - Achieved real-time performance but faced limitations in detecting small objects and precise localization.
- YOLOv2
  - Incorporated anchor boxes for improved bounding box prediction accuracy.
  - Enhanced network architecture and training techniques to achieve better overall performance.

- YOLOv3
  - Further advancements in accuracy and speed.
  - Utilized the Darknet53 backbone for efficient feature extraction.
- YOLOv4
  - Introduced the CSPDarknet53 backbone, enhancing both speed and accuracy.
  - Implemented modifications in network design and loss functions.
- YOLOv5
  - Provided scaled versions (YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x) catering to specific applications and hardware requirements.
  - Open source and actively maintained by Ultralytics, emphasizing ease of use, training, and deployment.
- YOLOv6
  - Continued improvements in accuracy and speed.
  - Introduced enhancements in the network architecture and training techniques.
- YOLOv7
  - Expanded support for multiple vision tasks, including object detection, segmentation, pose estimation, tracking, and classification.
- YOLOv8
  - Utilized a similar backbone as YOLOv5, with changes in the CSPLayer, now called the C2f module, to enhance detection accuracy.
  - Supported multiple vision tasks such as object detection, segmentation, pose estimation, tracking, and classification.
- YOLOv9
  - Uses Programmable Gradient Information (PGI) to improve learning from data.
  - Designed to minimize information loss during processing for better results.

This evolutionary reflects the commitment of the YOLO framework to advancing network design, backbone architecture, and training techniques. Notably, the transition to different frameworks, such as PyTorch and PaddlePaddle, has further influenced the development and performance of YOLO versions.

In summary, YOLO stands as a groundbreaking object detection system, revolutionizing real-time computer vision applications with its speed, accuracy, and efficiency. Its continual evolution and adaptability across various domains underscore its significance in the field of object detection and recognition.

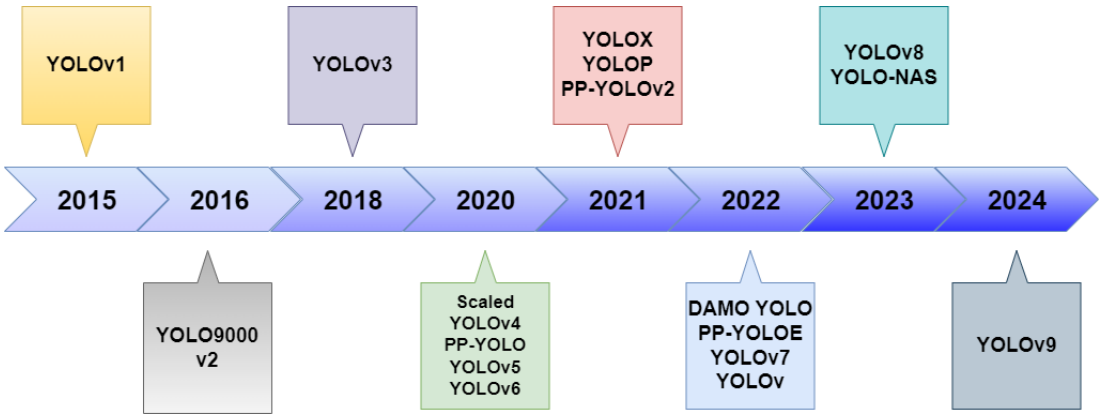


FIGURE 2.15: A timeline of YOLO versions.

2.4.6.5 Single Shot Multibox Detector (SSD)

The Single Shot Multibox Detector (SSD) algorithm is another influential framework in object detection. SSD employs a series of convolutional layers to predict object categories and bounding box offsets. Its ability to handle objects of varying sizes and scales makes it a valuable asset for tasks like lip-reading, where precise localization of lip movements is crucial.

The subsequent section explores the theoretical foundations of audio processing and its integration with computer vision, elucidating the synergistic approach employed in visual speech recognition.

2.5 Object Tracking Algorithms

Object tracking plays a pivotal role in computer vision applications, enabling the continuous monitoring and analysis of objects' movements over time. Various algorithms have been developed to address the challenges associated with tracking objects across frames in videos or image sequences. This section provides an overview of some prominent object tracking algorithms, highlighting their key characteristics and applications.

- **Traditional Statistical Methods:**
  - **Mean Shift Algorithm:** A non-parametric method that iteratively shifts the tracking window towards the mode of the probability distribution within the window. Particularly effective for tracking objects with consistent and distinctive color features.
  - **Kalman Filter:** Recursive, statistical estimation algorithms widely used for object tracking. They predict the next state of an object based on the previous state and incorporate measurements to refine predictions. Robust in handling noisy data.
  - **Correlation Filters:** Utilize the correlation between the object template and the image to track objects. Efficient in handling scale and translation variations, often used for real-time tracking applications.

  - **Multiple Hypothesis Tracking (MHT):** A probabilistic framework that considers multiple hypotheses for object tracks. Excels in situations where occlusions and uncertainties are prevalent, providing a robust approach to handle complex tracking scenarios.
  - **Particle Filter:** Represent the object's state with a set of particles, each evolving independently based on motion and observation models. Suitable for tracking objects with nonlinear and non-Gaussian dynamics.

- **Feature-based Approaches:**
  - **KLT Tracker (Kanade-Lucas-Tomasi):** Identifies key points in an image and tracks them across frames. Known for its feature-based approach, making it suitable for tracking objects with distinct visual features.
  - **Feature Point Matching:** Involves identifying and tracking specific feature points or keypoints in an image across frames. Techniques like SIFT (Scale-Invariant Feature Transform) and SURF (Speeded-Up Robust Features) fall under this category.

- **Deep Learning-based Approaches:**
  - **Deep SORT (Simple Online and Realtime Tracking with a Deep Association Metric):** Combines deep learning features with the SORT algorithm for object tracking. Leverages a deep neural network for feature extraction and employs the Hungarian algorithm for data association, achieving high accuracy in crowded scenes.
  - **Long Short-Term Memory (LSTM) Networks:** A type of recurrent neural network (RNN) applied to capture temporal dependencies in object tracking. Effective in scenarios where object motion patterns are crucial for accurate tracking.
  - **Siamese Networks:** Learn a similarity metric between object patches, enabling robust object tracking. Adept at handling variations in object appearance and often used in tracking-by-detection frameworks.

- **Comparison and Selection:**

  The choice of object tracking algorithm depends on the specific requirements of the application, as well as the characteristics of the tracked objects, such as lip movements in the context of visual speech recognition. Each algorithm has its strengths and limitations, and the selection process should consider factors like computational efficiency, robustness, and adaptability to different lip movement patterns.

Object tracking algorithms vary in their approaches, ranging from traditional statistical methods to state-of-the-art deep learning techniques. The choice of algorithm depends on the specific requirements of the tracking task, such as the nature of the tracked objects, environmental conditions, and real-time constraints. Researchers continue to explore novel algorithms and hybrid approaches to enhance the accuracy and efficiency of object tracking in diverse applications.

2.6 Conclusion

By establishing a strong theoretical foundation in computer vision, deep learning, object detection, and tracking, we equip ourselves with the necessary knowledge to develop and advance automated lip-reading systems with integrated object tracking capabilities. The insights gained from these concepts pave the way for the innovative techniques and methodologies detailed in the subsequent chapters of this thesis.

Chapter 3

Literature Review

3.1 Introduction

In the domain of Visual Speech Recognition (VSR), researchers have pursued two primary avenues to propel the advancement of automated systems. One track emphasizes constructing extensive databases, while the other track centers on innovating approaches for automated VSR system development.

3.2 Datasets Building

In the last two decades, there has been a remarkable upsurge in interest directed towards the development of visual speech recognition systems. Nonetheless, researchers in this scientific domain find themselves navigating along two intertwined paths. The first involves the procurement, compilation, and construction of datasets, while the second revolves around the proposition of efficacious methodologies and approaches.

The latter track inherently depends on the availability of diverse datasets that aptly reflect the complexities of real-world scenarios, encompassing variations in language, dialect, and lighting conditions, as well as accounting for diverse factors such as gender, age, and more. Researchers, in response to this imperative, have proactively put forth datasets for consideration. These datasets encompass both those meticulously crafted within controlled laboratory settings and those extracted from televised programs, each bearing its unique set of merits and demerits.

These datasets exhibit variances in dimensions such as size, linguistic diversity, the number of speakers, video quality, and other nuanced aspects. In delving deeper into this exploration, we aim not only to shed light on the existing datasets but also to categorize them based on their intended purposes. This categorization may include distinctions between datasets designed for characters/digits, words, and more extensive linguistic units such as sentences and phrases. Such a nuanced understanding of the landscape of datasets is integral for researchers striving to advance the field of visual speech recognition.

3.2.1 Content-Based Categorization

Categorize datasets based on the content they cover, including categories such as digits/letters, words, sentences, etc.

3.2.2 Availability-Based Categorization

Individual Letters and Digits This section lays the foundation for VSR by focusing on the accurate recognition of individual letters or digits. Datasets in this category include: **(i)** `AVletters` dataset, introduced by Matthews and Baker in 1998 [60], is dedicated to recognizing isolated letters and serves as a fundamental resource for understanding basic visual speech patterns. Originating from the University of East Anglia and the University of Manchester, it was the first audio-visual dataset and initially comprised 780 utterances from 10 speakers, each making three independent statements of 26 English letters. Due to shooting limitations in the earlier years, the video resolution is low at 376 pixels by 288 pixels, with 25 frames per second. After manually locating lip positions in each image, the entire image was cropped to 80 pixels by 60 pixels to form the final dataset. **(ii)** The `XM2VTS` dataset [62], a substantial multi-speaker database comprising 295 participants, is notable for digit recognition. Specifically designed for personal identification, each subject in the database was requested to articulate two continuous digit strings and one phonetically balanced sentence, and **(iii)** The `CUAVE` dataset presented by Patterson et al. [72] despite its seemingly modest speaker count compared to giants like XM2VTS and VALID. Its true power lies not in numbers, but in depth. Each of CUAVE's 36 participants contributes a staggering 260 utterances, meticulously choreographed into single and dual-speaker sessions. This dance of articulation unveils itself in individual speakers confidently enunciating 50 isolated digits, gracefully turning for profile recordings, and culminating in tackling 60 connected digits head-on. The complexity doesn't stop there. Dual sessions add another layer, capturing the ballet of two speakers, one voicing connected digits while the other waits silently, taking turns to repeat the sequence. This symphony of utterances, mirroring the nuances of real-world challenges like varying speaker positions and turn-taking, is precisely why CUAVE remains the undisputed maestro of ALR digit recognition training.

Words Moving beyond individual characters, this section focuses on recognizing complete words and phrases, providing datasets such as: **(i)** `The MIRACL-VC` dataset [84] emphasizes word-level recognition, enabling researchers to explore the challenges associated with context and word boundaries, the major highlight of this dataset is that it also contains depth images of the speakers, as illustrated in **Figure 3.3**, which adds to the training of any DL model. The dataset consists of 10 words and 10 phrases, with 15 speakers (5 men and 10 women) each uttering the words and phrases ten times each, totalling 3000 utterances. Every utterance in the sequence is an image of size 640 x 480 pixels with 25fps, and **(ii)** The `MODALITY` dataset [21] further advances this section by encompassing a diverse vocabulary, allowing for the study of VSR in a broader linguistic context.

Sentences and Phrases This advanced section addresses the complexities of recognizing sentences and phrases, which often involve coarticulation and context-dependent phoneme variations. Datasets within this category include: **(i)** `VIDTIMIT` [89] database comprises video and audio recordings of 43 volunteers (19 female, 24 male) reciting short sentences. Recorded in three sessions with intervals to capture changes in voice, appearance, and mood, it incorporates attributes relevant to verification system deployment. The sentences, drawn from the NTIMIT corpus, include 10 per person, with no repetition across training and test sections. The database features an extended head rotation sequence and is suitable for developing text-independent verification systems. It was recorded in a noisy office

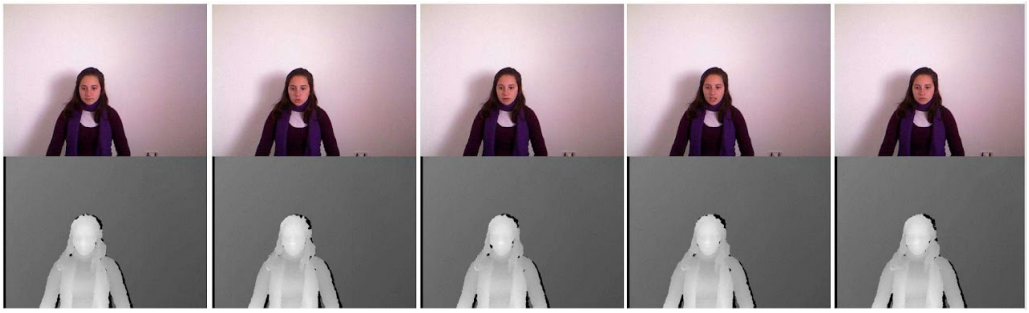


FIGURE 3.1: Frames from the MIRACL-VC1 dataset.

environment with a PAL digital video camera, resulting in a 3.5 Gb database. **(ii)** `OuluVS` In the study by Zhao et al. [121], the scope of investigation expands to encompass a broader range of sentences, exploring the impact of sentence length and complexity on Visual Speech Recognition (VSR). The OuluVS database, introduced in 2009 by a team from the University of Oulu in Finland, comprises video and audio data from 20 subjects uttering 10 phrases, each repeated five times. Additionally, the dataset includes videos capturing head movements from front to left, front to right, and without talking, repeated five times each. The OuluVS dataset serves as a standardized benchmark for assessing audiovisual speech recognition systems, featuring 20 participants delivering a total of 1000 utterances for the specified daily greeting phrases. This dataset marks a significant contribution to advancing lip-reading and automatic speech recognition systems. **(iii)** The `LRS` (Large-scale Multimodal Lipreading Dataset) [14] offers a comprehensive dataset for sentence and phrase-level lipreading, designed for large-scale studies in this domain. several other datasets incorporating the multidimensional data concept have been proposed, such as the Lip Reading Sentences 2 **LRS2** dataset [93]. Predominantly sourced from BBC news and TV shows, LRS2 features sentences with a maximum length of 100 characters, divided into training, validation, and test sets based on broadcast dates to ensure data dissimilarity. Noteworthy for its complexity, LRS2 encompasses diverse head poses and lacks speaker labels among its thousands of speakers, with clips captured at 160 × 160 pixels resolution and 25fps, See **Figure 3.3**.

FIGURE 3.2: **Top:** images from the BBC lip reading dataset. **Bottom:** mouth detection and tracking.

Another dataset, **LRS3-TED** [3], caters to visual and audio-Visual Speech Recognition (VSR) needs, comprising 5594 English TED and TEDx talks from YouTube, totalling over 400 hours of video. Face recordings, provided in .mp4 format, employ the h264 codec with a resolution of 224 × 224 pixels at 25 fps and are categorized into pre-train, train-Val,

TABLE 3.1: Some previous datasets.

Task	name	Language	classes	speakers	utterances	resolutions
Alphabet	**AVLetters** [60]	English	26	10	780	376x288,25fps
Alphabet	**AV@CAR** [70]	Spanish	26	20	800	768x576, 25fps
Digits	**XM2VTS** [62]	English	10	295	885	720x576, 25fps
Digits	**CUAVE** [72]	English	10	36	7000	720x480, 30fps
Words	**MIRACL-VC** [84]	English	10	15	1500	640x480, 15fps
Words	**MODALITY** [21]	English	182	35	231	1920x1080, 100fps
Words	**LRW** [12]	English	500	1000	400	256x256, 25fps
Words	**AVSD** [27]	Arabic	10	22	1100	1920x1080, 30fps
Words	**GRID** [16]	English	51	34	34000	720x576, 25fps
Phrases	**OuluVS** [121]	English	10	20	1000	720x576, 25fps
Sentences	**VIDTIMIT** [89]	English	346	45	430	512x384, 25fps
Phrases	**WAPUSK20** [105]	English	52	20	2000	640x480, 32fps
Sentences	**LRS** [14]	English	/	1000	118116	160x160, 25fps
Sentences	**CMLR** [14]	Mandarin	9	11	102076	64x128
Sentences	**GLips** [14]	German	500	100	250000	256x256
Sentences	**CN-CVS/Speech** [14]	Mandarin	75	2529	193,329	640x480
Sentences	**CN-CVS/News** [14]	Mandarin	465	28	13016	640x480
Sentences	**RUSAVIC** [14]	Russian	72	20	14000	1920x1080
Sentences	**OLKAVS** [14]	Korean	> 100	1107	250000	1920x1080

and test sets. Expanding upon these sections and datasets allows for a more nuanced exploration of VSR challenges and capabilities. Researchers can choose datasets that align with their specific objectives, whether it be character-level recognition, word-level understanding, or the complexities of sentences and phrases.

These datasets collectively contribute to the advancement of VSR technology across various linguistic contexts and applications. In summary, according to the dataset used, previous works can be roughly classified into three categories: *letters and digits*, *words*, and *phrases and sentences*. Our work falls in the third category.

3.3 Front-end: Lip Detection Approaches

Lip detection has recently become a focal point in the computer vision community, driven by its significance in various applications where visual information plays a vital role in enhancing system performance and resilience. Applications such as audio-visual speech recognition, lip synchronization, synthetic talking faces, and facial expression analysis all underscore the importance of accurate lip detection. Despite this, achieving precision and robustness in lip detection proves to be a formidable challenge, given the substantial variations arising from factors like the highly deformable nature of lips, diverse lip color tones, varying illumination conditions, and the presence of elements such as teeth, tongue, and facial hair.

In this section, we explore two main distinct categories of methods that contribute to the identification of the lip region in the context of visual speech analysis. These categories encompass Pixel Information-Based Methods, which focus on leveraging color differences for precise lip detection; and Model-Based Methods, exemplified by the Active Shape Model (ASM), utilizing statistical modeling to incorporate prior knowledge of lip shape. By delving into the nuances of these methodologies, we aim to establish a comprehensive

understanding of the diverse strategies employed in the field of lip detection, laying the groundwork for enhanced visual speech recognition systems.

3.3.1 Pixel-Centric Approaches:

pixel-centric approaches play a pivotal role in analyzing visual information at the pixel level. These methodologies focus on leveraging individual pixel characteristics, such as color, intensity, and texture, to extract meaningful insights and facilitate various tasks. This section explores pixel-centric approaches, shedding light on their applications, principles, and contributions within the domain of computer vision.

RGB approach: The RGB approach in lipreading involves leveraging color information encoded in the Red, Green, and Blue (RGB) channels of video frames to enhance lip movement analysis. Unlike grayscale approaches that focus solely on intensity variations, the RGB approach considers the distinct color information present in lip regions. This can provide additional discriminative features for lipreading systems, capturing nuances in lip appearance and aiding in the recognition of spoken words.

Chengjia Yang et al. [115] proposed an enhanced iteration of the region growing algorithm that capitalizes on the RGB colorspace for lip contour extraction. The method improves upon its predecessor by automatically selecting seed pixels through the estimation of color associations with adjacent pixels. After a sequence of image processing operations, the lip contour region is extracted using the watershed algorithm. Experimental analysis underscores the method's simplicity and efficacy, achieving a notable accuracy rate of 80

Fuzzy Clustering: Fuzzy clustering (FC) in lip reading involves applying FC algorithms to group visual features extracted from lip movements, contributing to the identification and interpretation of spoken language [51, 49, 110, 86]. FC has gained popularity in lip segmentation owing to its capacity to precisely categorize pixels as either part of the lip or non-lip region, even in scenarios where the lip area is partially concealed or merges with the surrounding skin. Additionally, its capability to manage variations in lighting and noise renders FC a pragmatic option for lip segmentation in real-world settings.

3.3.2 Model-Based Methods:

Active Contour Models (ACM): Kass, Witkin, and Terzopoulos [44] introduced Active Contour Models (ACM), or snakes, in 1988, revolutionizing image segmentation. Their method initiates a deformable contour near the object, iteratively adjusting it to minimize an energy function, balancing internal smoothness and external feature attraction forces. This approach provides a flexible and adaptive tool for object delineation in computer vision.

Delmas et al [23] introduce an innovative method for extracting resilient lip boundaries using active contours, commonly known as snakes. The paper addresses key challenges, including initialization, parameter estimation, and convergence of the snake algorithm. To enhance snake performance, the authors propose a new approach for extracting the mouth region of interest (ROI) by leveraging gray level image and gradient information. This method employs spatially varying coefficients to guide the snake's evolution, ensuring it maintains a mouth-like shape, demonstrating robustness in accommodating speaker changes. The technique holds promise for applications in video-conferencing and various

realms of image processing and computer vision. Other studies have also explored aspects of lip localization, such as the works by Barnard et al. [8] and Eveno et al. [28], contributing valuable insights to the field.

Despite the promise of snakes in lip localization, challenges arise, such as misfitting to features like the nose or chin and difficulty handling sharp curves and facial hair. Tuning parameters for optimal performance can be a time-consuming process, taking several seconds.

Active Shape Models: ASMs are deformable models used for shape analysis and object recognition in computer vision. Introduced by Cootes et al. [18], ASMs combine shape information with appearance statistics, adapting to the variability within a training dataset. Widely applied in tasks like facial landmark detection and medical organ segmentation, ASMs exhibit adaptability and robustness in diverse computer vision applications.

This work [55] focuses on continuous audio-visual speech recognition and prominently features the use of an appearance-based model (ASM) for robust lip tracking and visual speech feature extraction. The ASM is utilized to learn linguistically important features from example images, enabling the accurate localization, tracking, and recovery of visual speech information.

Dinh and Milgram [65] introduced a multi-feature ASM that integrates the normal contour, gray block, and Gabor wavelet for lip localization. This approach addresses the challenge faced by single-feature ASMs, which tend to converge to local minimum values in noisy environments, including factors like beard, wrinkles, lip color variations, and low skin color contrast. Some other work that have been done on ASM [63, 58, 56]

Active Appearance Models: Rothkrantz [87] employed Active Appearance Models (AAM) for lip tracking, leveraging the model's ability to capture both shape and texture changes. AAM establishes a statistical model for shape and a global gray model for texture, demonstrating improved lip detection results in controlled conditions. However, challenges arise in natural conditions, where internal factors like facial features (e.g., beard, wrinkles) and external factors like changes in light and background introduce interference, impacting the accuracy of lip detection. Unlike Active Shape Models (ASMs), AAMs extend beyond edge profiles along landmarks by incorporating the entire grayscale information of the image along with shape details. This extension enhances the model's representation capabilities, making AAMs valuable for applications in facial analysis and recognition.

Some other work that have been done on AAM [6, 17, 60]

In conclusion, the study of front-end lip detection highlights its critical role in various computer vision applications. Despite its significance, challenges persist due to the complex nature of lips, encompassing deformability, diverse color tones, and varying lighting conditions.

This section explores Pixel-Centric Approaches, such as the RGB approach and fuzzy clustering, focusing on pixel-level analysis for enhanced lip movement interpretation. Additionally, Model-Based Methods, including Active Contour Models (ACM), Active Shape Models (ASM), and Active Appearance Models (AAM), employ statistical models for shape and appearance variations in lip detection.

While these approaches offer unique advantages, challenges like misfitting to facial features and time-consuming parameter tuning underscore the complexity of achieving precision in lip detection. In Chapter 6, we will contribute to this field by proposing a

deep learning approach (YOLO) for lip detection and tracking, aiming to enhance the effectiveness of visual speech recognition systems.

3.4 Back-end: VSR Systems Approaches

The second track of research in Visual Speech Recognition (VSR) is dedicated to introducing innovative approaches aimed at enhancing automatic recognition systems. Central to these systems is the task of extracting features from the mouth region in images, a process laden with challenges such as coping with diverse lighting conditions and accommodating color variations. The proficiency of these systems critically depends on their capacity to precisely classify the features extracted from visual data.

This research trajectory can be bifurcated into two distinct yet interconnected parts:

3.4.1 Traditional Machine Learning Approaches:

Even though Visual Speech Recognition (VSR) isn't new, it became really interesting in the 1990s because of improvements in Computer Vision (CV) and Machine Learning (ML). During that time, scientists came up with many ways to look at videos of people talking and understand what they're saying just by seeing them. These early studies laid the foundation for how we do VSR today and keep inspiring new research in the field.

In the early VSRs, features were extracted using handcrafted methods. These are like specific techniques and rules designed by researchers to find important information in the videos. Even though we now have more advanced technologies, looking back at these older methods helps us understand how VSR has evolved.

In this study, we looked closely at fifteen important works in traditional VSR. We wanted to see how they're similar to today's VSR and understand the methods they used, especially those handcrafted methods for feature extraction. This review helps us learn about the progress and changes in VSR research, providing useful information for researchers and practitioners in the world of traditional Visual Speech Recognition.

The selection of robust features is paramount for the effectiveness of VSR systems. Among the prominent features are the Active Appearance Model (AAM) [20, 38], which encapsulates facial appearance variations, the Discrete Cosine Transform (DCT) [54, 91], employed for capturing signal energy compaction, and the Histogram of Oriented Gradients (HOG) [84], known for its ability to represent local object shape and texture. These features, crucial components of visual-speech systems, are typically determined by domain experts, leveraging their insights to encapsulate the nuances of speech-related visual cues. To further enhance the discriminative power of these features, machine learning techniques come into play. Classification tasks are commonly executed using advanced methodologies such as Hidden Markov Models (HMM) [48], a staple in temporal modeling, and Support Vector Machines (SVM) [124], renowned for their proficiency in pattern recognition. The synergy between carefully selected features and sophisticated classification techniques forms the foundation for robust and accurate ALR systems, paving the way for advancements in the field.

3.4.2 Deep Learning Approaches

In recent years, a transformative wave has swept through the field of Visual Speech Recognition (VSR), ushering in a new era dominated by Deep Neural Network (DNN)-based systems. This paradigm shift is marked by the rapid displacement of each constituent block in traditional systems, ultimately paving the way for the ascendancy of end-to-end DNN architectures. Our comprehensive research navigates this evolving landscape, focusing on the most prevalent DNN architectures within the domain of Audio-Visual Speech Recognition (ALR) systems. Delving into the these architectures, our exploration centers on the diverse variations sharing a common baseline structure – the fusion of Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs).

This shift harnesses DL approaches and often employs end-to-end models. This part has evolved through three distinct stages, each reflecting the progressive integration and refinement of deep learning methodologies in Visual Speech Recognition (VSR).

Transition to CNNs for Feature Extraction: - In the initial stage, [69] combined CNN deep learning methods and HMM, and applied it to lip reading problems. In this combination, CNN was used for feature extraction and HMM was used for classification.

This study [68] introduced an AVSR system employing deep learning for audio and visual features, combined using a Multistream Hidden Markov Model (MSHMM) for isolated word recognition. The denoising autoencoder enhanced audio feature robustness, while the CNN improved visual feature prediction. The MSHMM demonstrated reliable AVSR performance by adapting to changing signal reliability. Future work will explore real-world applications and assess the approach's robustness in dynamic environments.

This hybrid approach, observed in the outlined study [59], delves into the synergies between DL and HMM. Employing a 3D-2D-CNN-BLSTM architecture with an end-to-end CTC-based training methodology, the study contrasts this with the traditional BLSTM-HMM hybrid model. Bottleneck features extracted from the DL architecture serve as a common ground for both training paradigms. The exploration aims to unravel the comparative performance of end-to-end training against the more conventional hybrid model, providing valuable insights into their respective merits and limitations.

- This marked a significant departure from manual feature engineering, highlighting the potential of DL in automating the extraction of discriminative visual features for speech recognition.

CNN-Based Approaches for Feature Extraction and Classification: - Building on the success of the first stage, the second stage witnessed a pronounced shift towards employing CNN-based approaches for both feature extraction and subsequent classification.

Chung et al. in [14] exclusively utilized CNN for both feature extraction and classification. The study formulates a multi-way classification problem using architectures based on image classification models, particularly the VGG-M model. Four models, inspired by human action classification, are developed and compared. They differ in how they process input frames, maintaining VGG-M's configuration for direct performance comparison. The architectures include 3D Convolution with Early Fusion (EF-3), 3D Convolution with Multiple Towers (MT-3), Early Fusion (EF), and Multiple Towers (MT). See **Figure 3.3**.

- Researchers recognized and harnessed the substantial advantages offered by DL, demonstrating the efficacy of end-to-end CNN architectures in enhancing the accuracy and efficiency of VSR systems.

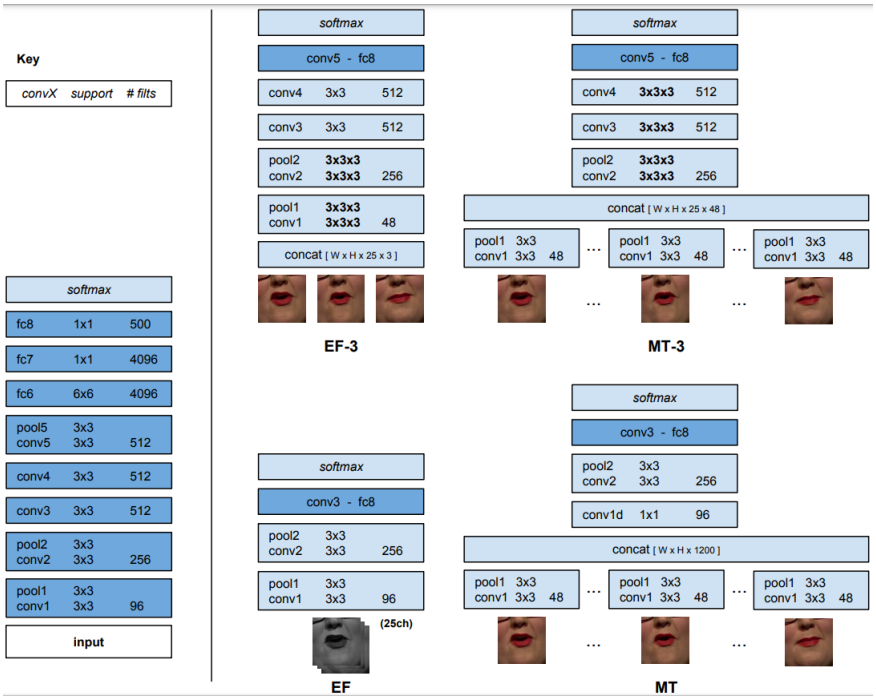


FIGURE 3.3: The overview of the work [14] shows the use of CNN in feature extraction and classification.

Integration of CNN with RNN: One of the main challenges in VSR is to capture the temporal dependencies and long-term dependencies in the visual speech signals. Traditional recurrent neural networks (RNNs), such as convolutional neural networks (CNNs), are not suitable for this task because they suffer from vanishing or exploding gradients, which limit their ability to learn long-term dependencies. Moreover, RNNs have difficulty in maintaining long-term memory and forgetting irrelevant information.

To overcome these limitations, long short-term memory (LSTM) and gated recurrent unit (GRU) networks have been proposed as alternative RNN architectures. LSTM networks have a special structure that allows them to store and access information over long time intervals without forgetting or leaking information. GRU networks are simpler versions of LSTM networks that use gates to control the flow of information within each cell. Both LSTM and GRU networks have shown superior performance on various VSR tasks compared to CNNs.

Garg et al. [32] proposed a method involving the utilization of a pre-trained VGGNet to extract image features. This set of features was then fed through multiple LSTM layers to capture temporal dependencies and derive the final classification label from the model's output. To address concerns related to overfitting and gradient instability, the authors incorporated batch normalization and dropout as regularization techniques, while also implementing gradient clipping. Despite these efforts, the LSTM-based models failed to surpass alternative methods, citing challenges such as prolonged training times and the necessity for recurrent networks throughout the model for effective feature extraction with temporal context. However, it is noteworthy that recent advancements in lip reading using LSTMs show promise, indicating their potential to outperform traditional methods.

The research by Rajneel et al. [25] explores three CNN models for lip reading: a basic model, an enhanced version with LSTM, and a GRU-integrated model, showing promising results. Despite initial challenges with a small dataset, hyperparameter tuning,

especially for GRU, improved accuracy. The study highlights the potential of volumetric convolutions and utilizes the MIRACL-VC1 dataset, demonstrating the efficacy of the proposed CNN models with depth and color pictures. This aligns with the growing interest in deep learning for visual speech recognition, showcasing the potential of multi-view silent videos for improved speech reconstruction. The use of benchmark datasets underscores the practical relevance in various fields, including visual speech recognition, face detection, and biometrics.

Sarhan et al. [90] proposed a model consists of three stages: pre-processing, encoder, and decoder. In the pre-processing stage, the mouth area is extracted from the frames of movements using the dlib and OpenCV libraries. In the encoder stage, the model uses inception layers, gradient preservation layer, and bidirectional GRU layer to capture both forward and backward information flows. In the decoder stage, the attention layer is used to compute the attention weights, which determine how much importance should be given to each word in the source to determine the output sentence. The final output is calculated using the SoftMax function in the dense layer. The model also uses LSTM and GRU networks to capture temporal dependencies and improve the accuracy of the lip-reading system.

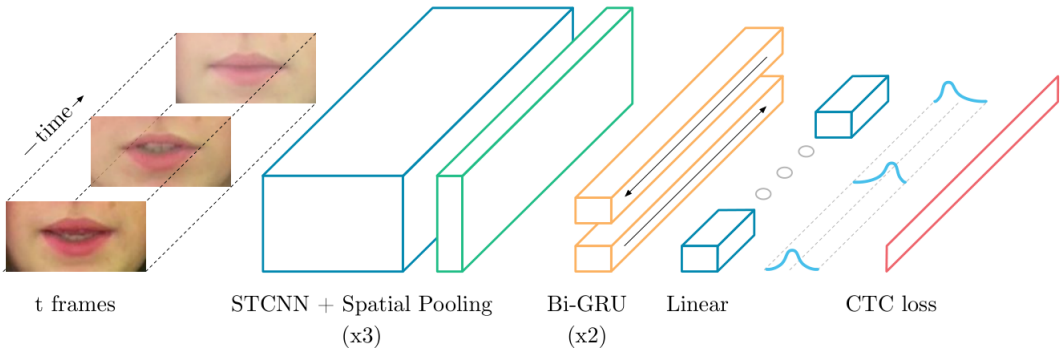


FIGURE 3.4: LipNet utilizes a sequence of T frames as input, processed by 3 layers of spatiotemporal CNNs, followed by spatial max-pooling. Extracted features undergo processing by 2 Bi-GRUs, with each GRU output time-step processed by a linear layer and softmax. The end-to-end model is trained with CTC.

Assael et al. [5] introduce LipNet, the pioneering end-to-end sentence-level lipreading model. Remarkably, LipNet achieves an impressive accuracy of 95.2% in lipreading on the GRID corpus, surpassing traditional approaches. The model leverages spatio-temporal convolutional neural networks (STCNNs) trained end-to-end, extracting spatio-temporal features from video frames. These features undergo processing by recurrent neural networks (RNNs) and the connectionist temporal classification loss (CTC), enabling LipNet to handle variable-length input and output sequences without explicit alignments. The crucial CTC loss eliminates the need for aligned training data, allowing the model to make sentence-level predictions Figure 3.4 illustrate the architecture of the LipNet. LipNet's success lies in its ability to focus on phonologically significant regions in videos, as evidenced by saliency visualizations. Notably, LipNet outperforms human lipreading baselines and previous state-of-the-art models

These studies demonstrate that LSTM and GRU networks are effective solutions for VSR problems. However, there is still room for improvement in terms of accuracy, robustness, scalability, and efficiency of these models (See Tables 3.2, 3.3, and 3.4).

TABLE 3.2: Deep learning architectures for lipreading (part 1).

year	Reference	Front-end	Back-ent	Database	Task	WRR(%)	
2011	Ngiam et al. [64]	Autoencoder + RBMs		AVLetters	Alphabet	64.40%	
				CUAVE		68.70%	
2014	Noda et al. [69]	CNN	GMM-HMM	ATR	Words	37.00%	
2015	Noda et al. [67]	CNN	MS-HMM	ATR	Words	22.50%	
2015	Sui et al. [99]	Autoencoder + DCT + LDA	HMM	AusTalk	Digits	67.80%	
2015	Sui et al. [98]	DBM+DCT+LDA	HMM	AusTalk	Digits	69.10%	
2015	Ninomiya et al. [66]	DBN	MS-HMM	CENSREC1-AV	Digits	39.30%	
2016	Petridis et al. [74]	DBNFs + DCT	LSTM	AVLetters	AlphabetPhrases	58.10%	
				OuluVS		81.80%	
2016	Hu et al. [40]	RTMRBM	SVM	AVLetters	Alphabet	64.63%	66.04%
				AVLetters2		31.21%	74.77%
2016	Takashima et al. [101]	CBNs	HMM	ART	words	52.00%	
2016	Garg et al. [32]	CFI+VGG	LSTM	MIRACL-VC1	Words + Phrases	76.00%	
2016	Wand et al. [106]	Feed Forward	LSTM	GRID	Sentences	79.60%	
2016	Li et al. [50]	CNN	HMM	ATR	Words	71.76%	
2016	Chung and Zisserman [13]	VGG-M	LSTM	OuluVS2	Digits/Phrases	31.90%	25.40%
		SyncNet				94.10%	92.80%
2016	Lee et al. [48]	CNN	LSTM	OuluVS2	Digits/Phrases	83.80%	
2016	Zisserman et al. [125]	PCAnn+LSTM	HMM	OuluVS2	Digits/Phrases	79.00%	
		MIN				86.10%	80.00%
2016	Saitoh et al. [88]	AlexNet		OuluVS2	Digits/Phrases	88.30%	80.90%
		GoogleNet				88.50%	81.90%
2016	Chung and Zisserman [12]	EF		LRW	Words	57.00%	
		MT				61.10%	
		EF-3				43.90%	
		MT-3				46.20%	
2016	Assael et al. [5]	3D CNN	Bi-LSTM+CTC	GRID	Sentences	95.20%	
2017	Rahmani and Almasganj [80]	DBNF	GMM-HMM	CUAVE	Digits	63.40%	
			DNN-HMM			64.90%	
2017	Gutierrez and Robert [36]	VGG-16	LSTM	MIRACL-VC1	Words + Phrases	79.00%	
2017	Wand et al. [107]	Feed-forward	LSTM	GRID	Sentences	83.30%	

TABLE 3.3: Deep learning architectures for lipreading (part 2).

year	Reference	Front-end	Back-ent	Database	Task	WRR(%)
2017	Petridis et al. [73]	Autoencoder + LSTM	Bi-LSTM	OuluVS2	Phrases	84.50%
				CUAVE	Digits	78.60%
2017	Petridis et al. [76]	Autoencoder + Bi-LSTM	Bi-LSTM	OuluVS2	Phrases	91.80% 98.60%
2017	Petridis et al. [78]	Autoencoder + Bi-LSTM	Bi-LSTM	OuluVS2	Phrases	96.90%
2017	Chung et al. [14]	CNN+LSTM	LSTM+ Attention	GRID	Sentences	97.00%
				LRW	Words	76.20%
				LRS2-BBC	Sentences	49.80% 86.10%
2017	Chung and Zisserman [11]	CNN+LSTM	LSTM + Attention	OuluVS2	Phrases	91.10%
				MV-LRS	Sentences	43.6%
2017	Torfi et al. [102]	3D CNN	Contrastive loss	LRW	Words	98.50%
2017	Stafylakis and Tzimiropoulos [96]	3D CNN + ResNet	Bi-LSTM	LRW	Words	83.00%
2017	Stafylakis and Tzimiropoulos [97]	3D CNN + ResNet + word boundaries	Bi-LSTM	LRW	Words	88.08%
2018	Wand et al. [108]	Feed-forward	LSTM	GRID	Sentences	84.70% 95.90%
2018	Petridis et al. [79]	Autoencoder	Bi-LSTM	AV Digits	Phrases	69.70%
					Digits	68.00%
2018	Chung and Zisserman [10]	CNN	LSTM-5	OuluVS2	Phrases	94.10%
				LRW	Words	71.50%
					Sentences	
2018	Hashmi et al. [37]		CFI + CNN + BN + Softmax	MIRACL-VC1	Words + Phrases	52.90%
2018	Fung and Mak [31]	3D CNN	Bi-LSTM	OuluVS2	Phrases	87.60%
2018	Shllingford et al. [92]	3D CNN + Bi-LSTM	FST	LSVSR	Sentences	59.10%
2018	Stafylakisa et al [95]	3D CNN + ResNet	Bi-LSTM	LRW	Words	88.08% 98.10%
2018	Petridis et al. [77]	3D CNN + ResNet	Bi-GRU	LRW	Words	82.00% 98.00%
2018	Petridis et al. [75]	3D CNN + ResNet-18 + Bi-LSTM (EF)	Bi-LSTM + Atttentio-CTC	LRS2-BBC	Sentences	57.90% 96.40%
		3D CNN + ResNet-18 + Bi-LSTM (LF)				57.90% 95.30%
2018	Afouras'et'al. [1]	3D CNN + ResNet	Bi-LSTM+CTC	LRS2-BBC	Sentences	37.80%
			FC+CTC			45.00%
			Transformer			50.00%
2018	Afouras'et'al. [2]	3D CNN + ResNet	TM-CTC	LRS2-BBC	Sentences	45.30% 91.80%
			TM-seq2seq			51.70% 91.50%
2018	Xu'et'al. [114]	3D CNN + HW2D CNN	BiGRU + Attention-CTC	GRID	Sentences	97.10%

TABLE 3.4: Deep learning architectures for lipreading (part 3).

year	Reference	Front-end	Back-ent	Database	Task	WRR(%)
2018	Yang et al. [116]	2D CNN DensNet3D 2D+3D CNN	Bi-GRU	LRW-1000	Words	25.76% 34.76% 38.19%
2019	Courtney and Sreenivas [19]	Res-BiConvLSTM		LRW	Words	85.20%
2019	Fenghour et al. [29]	Decoder-Encoder LSTM		TIMIT	Sentences	72.20%
2019	Jang et al. [43]	CFI + QVGG + Committee		OuluVS2	Phrases	90.90%
2019	Zhou et al. [123]	CNN	Bi-LSTM + MD_ATT_MC	Chinese TV OuluVS2	Sentences Phrases	93.15% 93.72%
2019	Mesbah et al. [61]	CFI + Hahn CNN		LRW GRID	Words Sentences	58.02% 98.70%
2019	Margam et al. [46]	3D CNN+2D CNN	Bi-LSTM-CTC	Indian English	Sentences	87.70%
2019	Weng and Kitani [111]	Two-Stream+I3D	Bi-LSTM	LRW	Words	84.11%
2019	Zhang et al. [119]	VGG-M + ResNet + Bi-LSTM+CTC	GRU-Attention	CCTV	Sentences	50.20%
2019	Wang [109]	ResNet-DenseNet3D	BiConvLSTM	LRW LRW-1000	Words	83.34% 36.91%
2020	Zhang et al. [120]	3D CNN + ResNet (cutout)	Bi-GRU	LRW LRW-1000	Words	85.2% 45.24%
2020	Xiao et al. [113]	flow: ResNetGray: 3DCNN + ResNet	Bi-GRU	LRW LRW-1000	Words	84.13% 41.93%
2020	Luo et al. [57]	3D CNN + ResNet (PCPG)	Bi-GRU	LRW LRW-1000	Words	83.50% 38.70%
2020	Zhao et al. [122]	3D CNN + ResNet (LMIM-GMIM)	Bi-GRU	LRW LRW-1000	Words	84.41% 38.79%

In conclusion, our survey provides a comprehensive and insightful overview of the dynamic landscape of DNN architectures in ALR. The myriad choices and considerations explored within this survey illuminate the evolution of visual speech recognition systems, offering valuable insights for researchers, practitioners, and enthusiasts invested in the cutting-edge advancements of this transformative field.

3.5 Arabic Visual speech Recognition

3.5.1 The Arabic Language

Despite being one of the oldest languages globally and currently ranking as the fifth most widely spoken language with approximately 350 million speakers, Arabic has received limited attention in research efforts [118]. Covering a vast geographical area, Arabic is a Semitic language and holds the status of one of the six official languages in the United Nations (UN). Remarkably, it serves as the first language for over 300 million native speakers, positioning it as the fourth most spoken language, following Mandarin, Spanish, and English [35]. Additionally, due to its role as the language of religious instruction in Islam, numerous individuals possess at least a basic knowledge of Arabic. It's essential to note that "Arabic" encompasses various dialects. Among these, Modern Standard Arabic (MSA) stands out as a formal standard utilized across all Arabic-speaking countries, representing a modernized version of Classical Arabic with an updated vocabulary [45].

3.5.2 Visemes: Fundamental Concepts

Visemes play a fundamental role in Visual Speech Recognition (VSR), commonly known as lipreading. Unlike phonemes, which represent basic sound units in spoken language, visemes correspond to the visual cues associated with specific speech sounds. In lipreading, visemes are the distinct visual representations of speech sounds, manifested through lip and facial movements. In lipreading, visemes are essential for capturing the visual distinctions between speech sounds that may not be distinguishable solely through auditory information. These visual cues encompass various aspects such as lip shapes, mouth movements, and facial expressions, all of which are crucial for comprehending spoken language, particularly in noisy or silent environments.

In the context of the Arabic language, researchers primarily focus on the challenges and techniques associated with viseme recognition within Modern Standard Arabic (MSA). The specific characteristics of Arabic pose unique challenges for VSR, as certain visemes may correspond to multiple phonemes, leading to ambiguity in interpreting spoken language solely through lipreading. Therefore, it's crucial to address these challenges and develop effective techniques for extracting nuanced Arabic lip movements through in-depth analysis.

Figure 3.5 illustrates common visemes in Arabic and their significance in the VSR process. Addressing these challenges requires advanced techniques and technologies, which we will discuss further in this paper. From this example presented in Figure 3.5, we can see that the phonemes "م" and "ب" share the same mouth shape (Viseme). As a consequence, differentiating between these two phonemes is a complex and task-prone process. Similarly, the phonemes "ر","ل","ض","د","ن","ث", present analogous challenges due to their shared viseme.

Previous research efforts have explored the classification of visemes, with a focus on the English language (e.g. [112, 30, 42]). However, the outcomes of these studies were not entirely satisfactory. Our study aims to go beyond mere visemes identification and recognition. Instead, we intend to develop DL models capable of being trained on complete sentences and automatically extracting relevant features.

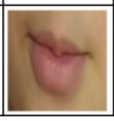
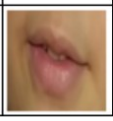
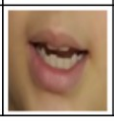
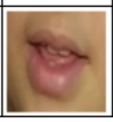
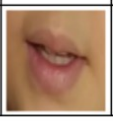
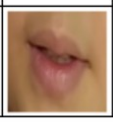
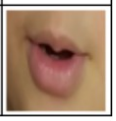
Visemes	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
Corresp. phonemes	ب /b/ م /m/	ف /f/	ت /t/ ن /n/ د /d/ ض /ð/ ر /r/ ل /l/	ط /tˤ/ ذ /ð/ ث /θ/	ص /sˤ/ ز /z/ س /s/	ج /dʒ/ ش /ʃ/	ي /j/	غ /ɣ/ ك /k/ خ /x/	ع /ʕ/ ق /q/ ح /ħ/ أ /ʔ/ ه /h/	و /w/
Shape										

FIGURE 3.5: Some visemes with their correspondence phonemes.

3.5.3 The Arabic Datasets And Approaches

The Arabic Visual Speech Dataset (AVSD) [27] is a dataset designed for visual speech recognition (VSR) in Arabic. It contains 1100 videos of 10 daily communication words spoken by 22 Arabic speakers. The videos were recorded using smartphone cameras in high resolution and high framerate. The dataset was taken under realistic conditions inside various rooms in different indoor illuminations. The dataset was designed to be useful for researching and developing visual speech recognition applications suitable for mobile devices in addition to other uses.

The study conducted by Dweik et al [26]. (2022) utilized a dataset comprising 1051 videos of 73 native Arabic speakers (33 females and 40 males). These videos featured the speakers uttering 10 common Arabic words in various environments. To ensure dataset generalization, the videos were captured using different cameras, at different distances, and with different background lighting settinThe dataset used in this study gs. The smartphones used for recording were placed in holders during the capture process. Each of the 73 speakers mostly uttered each of the 10 words once, although a few of them repeated the words between two and four times. This approach aimed to account for the diverse characteristics among individuals, including variations in speaking speed, mouth shape and movements, lips geometric features, tongue size (indicated by the redness of the mouth frame), alveolar ridge, teeth, braces, moustache, beard, and makeup.

The dataset used in the research [4] is a crucial aspect of the deep learning-based approach for Arabic visual speech recognition. Since no public Arabic datasets were available, the researchers collected their dataset to address this gap. The dataset comprises 2400 videos for Arabic digits and 960 videos for Arabic phrases, totalling 3360 videos. It was collected from 24 native speakers (14 males and 10 females), with each speaker pronouncing 10 Arabic numerical words and 4 phrases, and each number and phrase repeated 10 times. The resolution of each frame is 1920 × 1080, with approximately 25

frames per second, captured using a smartphone camera with a ring light stand. The data acquisition process took about one month and was conducted with white backgrounds, considering various factors such as background, speed of speech, speaker posture, and lighting.

In conclusion, while significant strides have been made in contributing Arabic audio-visual speech recognition datasets, challenges persist due to the limited availability of such resources. This scarcity emphasizes the importance of continued efforts in dataset creation to further advance research in this domain.

3.6 Our Work Positioning

In addressing the current landscape of Arabic lipreading systems, we find a notable dearth of research that systematically incorporates data building within existing approaches. Our claim is grounded in the conviction that the development of a pertinent dataset for Visual Speech Recognition (VSR) demands an augmentation of Arabic words and an expansion of videos capturing the same word to articulate nuanced behavioral information for speakers. While extensive efforts have been invested in building and adapting VSR for diverse languages, the existing body of work has overlooked the opportunity to leverage this wealth of experience for the benefit of Arabic languages.

The discerned gap in dedicated approaches becomes evident as we identify a deficiency in explicitly defining a dataset that not only captures event logs but also represents behavioral information for speakers in a manner conducive to the automatic interpretation of lip movements into written sentences. Previous endeavors have not sufficiently addressed the need for a structured dataset that aligns with the Arabic linguistic nuances. This oversight is significant, as Arabic, with its rich phonetic and morphological features, requires specialized attention to capture the subtleties of lip movements accurately.

Furthermore, our examination of prior work highlights the absence of a comprehensive framework designed to assist practitioners and researchers in harnessing existing frameworks as services for Arabic speech recognition systems. In response to these identified gaps, our work assumes a dual role. Firstly, we introduce a novel approach to Arabic Audio-Visual Speech that enriches the repertoire of automatic Arabic VSR systems. This approach is characterized by a methodical enhancement of dataset quality, ensuring that each video segment captures detailed and varied pronunciations of Arabic words. The inclusion of multiple speakers and diverse speaking conditions further amplifies the dataset's robustness, enabling the model to generalize better across different contexts.

Secondly, we present a framework that not only facilitates the incorporation of behavioral information into datasets but also guides practitioners and researchers in utilizing existing frameworks effectively for Arabic speech recognition systems. This framework is designed to be modular and adaptable, allowing seamless integration with current VSR technologies. By providing detailed guidelines on dataset creation, annotation, and preprocessing, our framework empowers researchers to build high-quality datasets that reflect the intricacies of Arabic lip movements.

Additionally, our work emphasizes the importance of multimodal data integration by proposing Audio-visual Arabic dataset, combining visual and audio data to enhance recognition accuracy and robustness. This integration leverages the complementary strengths of both modalities, leading to improved system performance. Model optimization is another critical focus, where we refine deep learning models to enhance real-time performance

and reliability under diverse conditions. Techniques such as hyperparameter tuning, architecture refinement, and transfer learning are employed to achieve this optimization.

Through these contributions, our work aims to propel advancements in Arabic VSR and bridge the existing gaps in both dataset construction and system implementation frameworks. By systematically addressing the challenges in Arabic VSR, we provide a comprehensive solution that not only advances the state of the art but also sets a new benchmark for future research in this domain. Our approach is poised to make significant strides in the field, fostering a deeper understanding and more accurate interpretation of Arabic lip movements in visual speech recognition systems.

3.7 Conclusion

In conclusion, our exploration of the related work in Visual Speech Recognition (VSR) reveals a dynamic landscape marked by two distinct tracks. The first track, rooted in traditional methods, underscores the historical evolution of VSR, where researchers initially relied on handcrafted features and well-established algorithms. This track highlights the foundational contributions of early studies in laying the groundwork for modern VSR systems.

Conversely, the second track introduces a paradigm shift, harnessing the power of Deep Learning (DL) methodologies. The three identified stages in this track reflect a progressive integration of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), showcasing the evolution towards end-to-end models. This contemporary approach capitalizes on the automated feature extraction capabilities of DL, promising enhanced accuracy and efficiency in VSR systems.

In the broader context, our examination of Arabic datasets and related works reveals a notable scarcity of dedicated research addressing the incorporation of data building within Arabic lipreading systems. This gap becomes particularly evident when compared to the wealth of research available for other languages. Our work seeks to fill this void by proposing a novel approach to Arabic Audio-Visual Speech, emphasizing the importance of an enriched dataset that captures behavioral information for speakers.

Furthermore, our positioning in the related workspace identifies a lack of explicit frameworks for the representation of behavioral information and a dearth of tools to facilitate the reuse of existing frameworks for Arabic speech recognition systems. To address these gaps, our work not only contributes a novel approach to Arabic VSR but also presents a comprehensive framework that guides practitioners and researchers in dataset creation and framework utilization.

As we delve into subsequent chapters, the synthesis of traditional and contemporary approaches, coupled with a dedicated focus on Arabic VSR, positions our work as a valuable contribution to the ongoing discourse in the field. Our efforts aim to propel advancements in the understanding of visual speech cues, ultimately enhancing the capabilities of automatic Arabic VSR systems.

Part III

Contributions

Chapter 4

BlidAVS10 Framework Overview

4.1 Introduction

In this chapter we provide an in-depth exploration of the BlidAVS10 framework, positioning it as an innovative solution in the realm of Visual Speech Recognition (VSR). This chapter meticulously unpacks the framework's functionality, focusing on its four pivotal services that collectively contribute to the VSR pipeline.

The first service revolves around the creation of a dataset, a foundational step crucial for subsequent analysis. This involves the generation of a set of videos primarily concentrated on scrutinizing individual phonemes or visemes. The dataset forms the cornerstone for further examination, comparison, and storage of instances under a standardized structure.

The second service delves into mouth region detection and tracking, a critical aspect of VSR. Utilizing advanced techniques such as the dlib library, YOLO, or any other technique, the framework excels in precisely identifying and extracting the mouth region within each frame of the videos generated in the dataset creation phase. This meticulous process ensures the accurate isolation of relevant visual cues.

Service three is dedicated to the customization of VSR models, providing developers with the flexibility to tailor attributes, CNN architecture, and hyper-parameters. This customization aspect is instrumental in aligning the VSR models with specific use cases, enhancing adaptability and performance.

The final service encapsulates the building, training, and evaluation of Deep Learning (DL) models. This phase ensures the implementation of the specified model configuration, its thorough training, and the subsequent evaluation of its performance. The integration of DL libraries, such as Pytorch, TensorFlow, and Keras, underscores the framework's commitment to leveraging cutting-edge technologies.

Throughout the chapter, a detailed narrative unfolds, elucidating the interplay between these services. Emphasis is placed on seamless communication and collaboration among the services, fostering a cohesive workflow. This interconnectedness ensures that each phase contributes harmoniously to the overarching goal of advancing Visual Speech Recognition.

In summary, This chapter provides a holistic view of the BlidAVS10 framework, elucidating its services and highlighting their synergistic collaboration. This comprehensive overview sets the stage for subsequent chapters, laying a solid foundation for the exploration and analysis of the framework's contributions to the field of Visual Speech Recognition.

4.2 Mind Map of the Proposed Approach

To facilitate understanding, we include a mind map (See Figure 4.1) that provides a comprehensive overview of our deep learning approach to VSR, outlining the key components

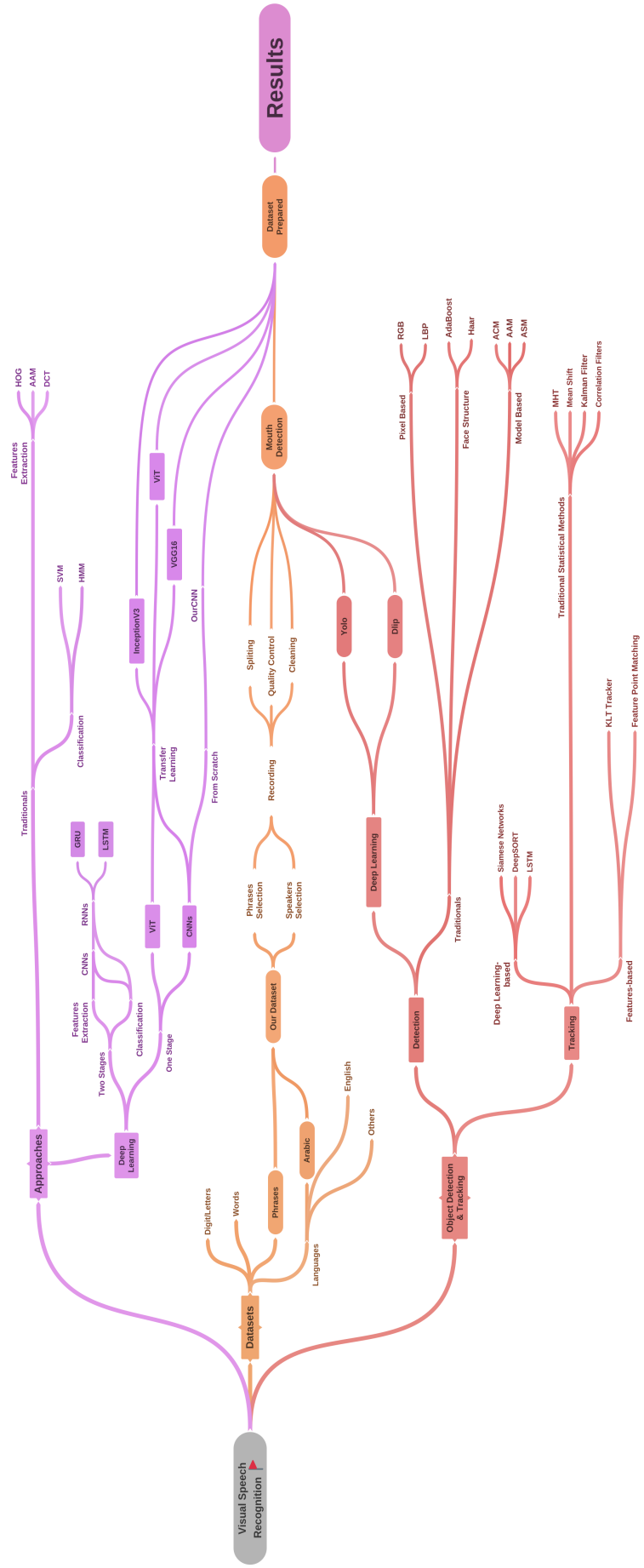


FIGURE 4.1: The roadmap of the BlidaAVS10

and methodologies employed throughout the research. This section delves into the various branches of the mindmap, elucidating the steps and processes involved in achieving robust and accurate VSR.

1. **Approaches** Our approach to VSR leverages advanced deep learning techniques and consolidates feature extraction and classification into a single model. We utilize ViT and CNNs with transfer learning from pre-trained models such as InceptionV3, VGG16, and ViT, while we have trained CustomCNN modified from scratch on both colored and grayscale images. This method ensures a streamlined and efficient recognition process, capable of capturing complex spatial hierarchies and temporal dynamics inherent in lip movements.

2. **Datasets** The dataset is a cornerstone of our VSR approach, emphasizing diversity and comprehensiveness. We have opted to create a new Arabic dataset that emphasizes phrases, with a significant subset dedicated to Arabic phrases alongside English and other languages. This dataset is crucial for training models that generalize well across different linguistic contexts.

    - *Phrases Selection*: Careful selection of phrases to cover a wide range of spoken content.
    - *Speakers Selection*: Inclusion of diverse speakers to enhance model generalization.
    - *Recording*: Systematic recording sessions ensuring high-quality data capture.
    - *Quality Control*: Rigorous quality control measures to maintain dataset integrity.
    - *Cleaning and Splitting*: Pre-processing steps to clean and split the data appropriately.

3. **Mouth Detection**

    Effective mouth detection is pivotal for accurate VSR, employing advanced deep learning methods. Our approach incorporates state-of-the-art object detection models to ensure high performance and precision. YOLO (You Only Look Once) is one such model, renowned for its real-time performance and accuracy, making it an excellent choice for mouth detection in dynamic video sequences. Additionally, we leverage Dlib, a traditional method celebrated for its reliability in facial landmark detection. While Dlib is rooted in classical techniques, our focus is on enhancing it with deep learning capabilities to improve detection robustness and adaptability. Using these advanced computer vision techniques provides a comprehensive and reliable solution for mouth detection in visual speech recognition systems.

4. **Results**

    The culmination of these efforts results in a well-prepared dataset, ready for further analysis and model training. The integration of various deep learning methods ensures a robust and versatile VSR system, capable of adapting to diverse scenarios and delivering high accuracy in visual speech recognition.

This detailed explanation provides a thorough understanding of the different components and methodologies depicted in the mindmap, highlighting the structured and comprehensive deep learning approach adopted in the research.

4.3 BlidAVS10 Mathematical Foundation

The "Mathematical Theoretical" section in Chapter 4 serves as a foundational exploration, providing a theoretical lens through which we scrutinize the fundamental concepts underlying the BlidAVS10 framework. As we embark on this theoretical, our objective is to unravel the mathematical details that form the bedrock of BlidAVS10's innovative approach to Visual Speech Recognition (VSR). By delving into the theoretical underpinnings, we aim to offer a deeper understanding of the relationships and mathematical frameworks that govern the creation of datasets and the meticulous workflow involved in building prediction models.

4.3.1 Dataset Concept

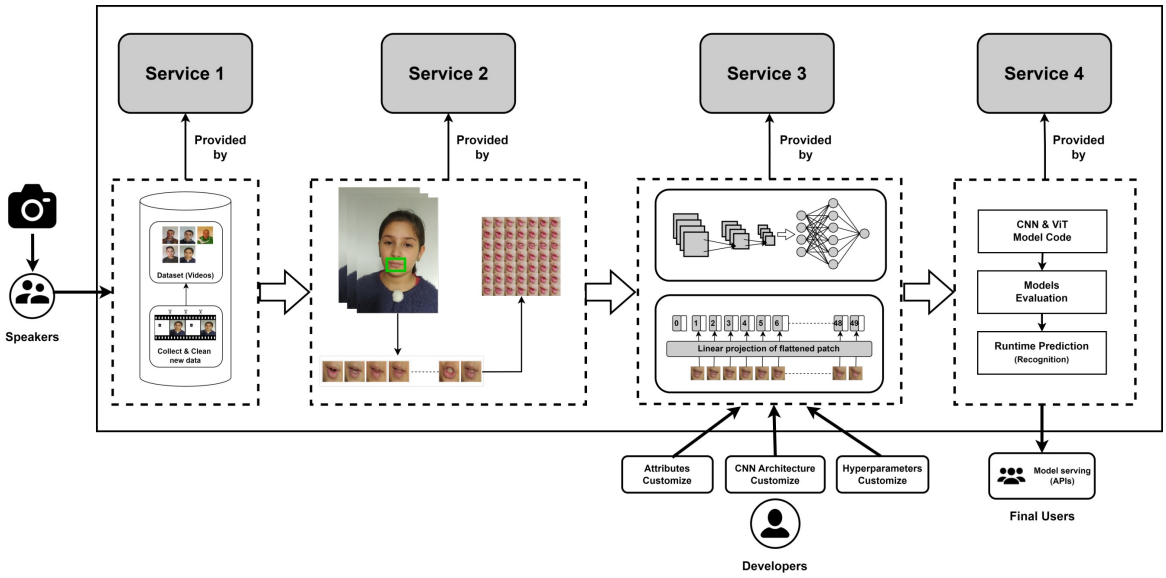


FIGURE 4.2: Overview of the `BlidaAVS10` Infrastructure

Various methods to tackle VSR have been suggested by researchers from both academia and industry. In the present day, artificial intelligence analysis plays a crucial role in enhancing VSR by utilizing the information stored in a trace model, which encompasses the system's data and event log.

The trace model representing features in dataset $\mathcal{D}$ is denoted as $\langle \mathsf{L_i}, \mathsf{V_L}, \mathsf{S_m}, \mathsf{G_A} \rangle$. This model establishes connections between the lip movement video $\mathsf{L_i}$ and an Arabic audio-visual representation defining lip movement visualization $\mathsf{V_L}$ (depicting lip movements), sequences of intermediate steps $\mathsf{S_m}$ (such as pronounced characters), and the goal analysis type $\mathsf{G_A}$ (e.g., phonetics, speech therapy, language instruction) tailored to specific requirements.

The Visual Speech Recognition (VSR) problem involves constructing a model $\mathcal{M}$ to generate voice or caption corresponding to the pertinent video $\mathcal{V}_x$ (see Figure 4.4). This model must adhere to a set of Quality of Service (QoS) constraints $\mathcal{C} = \{c_1, c_2, \ldots, c_t\}$, where each constraint c_t aligns with user's requirements or application constraints (e.g., real-time applications). Additionally, it must satisfy a set of QoS attributes $\mathsf{Attr}^{\mathsf{QoS}} = \{Attr_1, Attr_2, \ldots, Attr_m\}$ representing qualities such as minimal accuracy that the model $\mathcal{M}$ needs to fulfill.

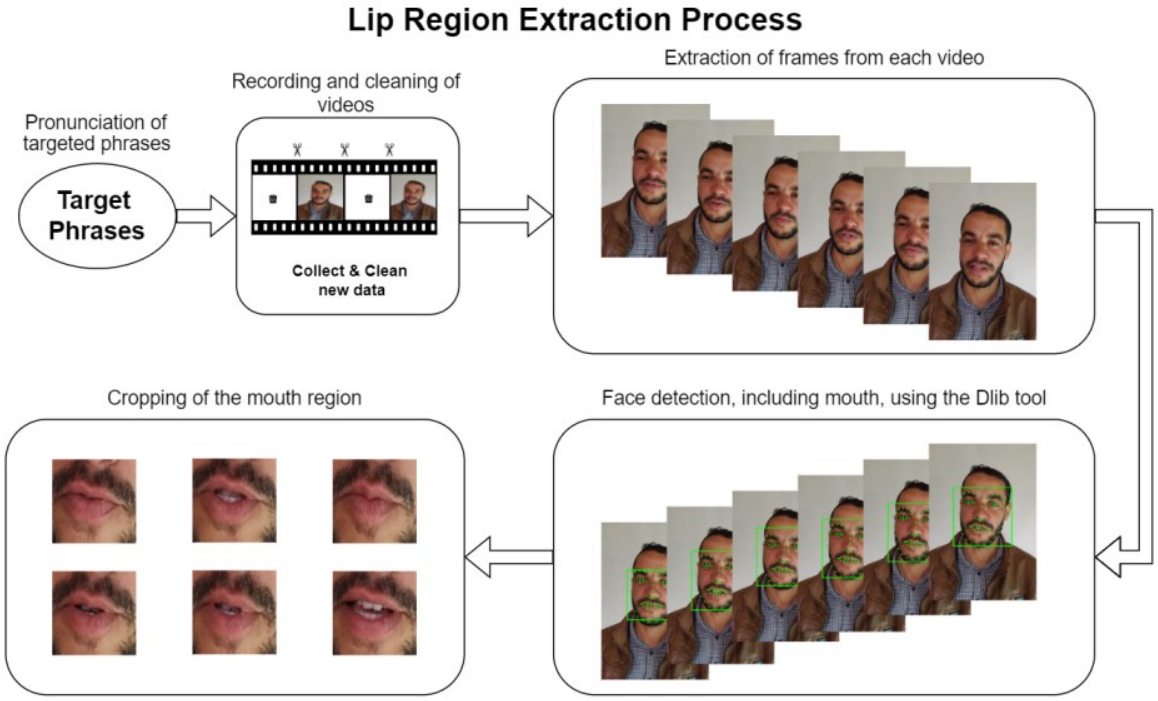


FIGURE 4.3: Illustration showcasing the exploration of various lip region extraction techniques

Building Dataset Phase:

- ***Input:*** Videos $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_n\}$
- ***Output:*** Dataset $\mathcal{D} = \langle \mathsf{L_i}, \mathsf{V_L}, \mathsf{S_m}, \mathsf{G_A} \rangle$

Building Model Phase:

- ***Input:*** Dataset $\mathcal{D} = \langle \mathsf{L_i}, \mathsf{V_L}, \mathsf{S_m}, \mathsf{G_A} \rangle$
- ***Constraint:*** $\mathcal{C} = \{c_1, c_2, \ldots, c_t\}$: users' requirements/application constraints

  e.g., real-time applications
- ***Goal:*** a set of attributes of QoS $\mathsf{Attr}^{\mathsf{QoS}} = \{\mathsf{Attr}_1, \mathsf{Attr}_2, \ldots, \mathsf{Attr}_m\}$ that the $\mathcal{M}$ has to satisfy (e.g., model's minimal accuracy).
- ***Output:*** VSR Models ($\mathcal{M}$)

Testing Model Phase:

- ***Input:*** a new video $\mathcal{V}_{new}$
  VSR Models ($\mathcal{M}$)
- ***Output:*** $\mathsf{Trans}_{\mathcal{V}_{\mathsf{new}}}$ produces the *voice* or *caption* of the relevant video

FIGURE 4.4: The VSR building process

Consider a given video, denoted as $\mathcal{V}_i$, which can be decomposed into a sequence of frames. Each set of visemes, represented as UV, corresponds to a subset of these

frames: $\langle \mathsf{UV}i_1, \mathsf{UV}i_2, \ldots, \mathsf{UV}i_n \rangle$. Furthermore, each individual visemes unit, such as $\mathsf{UV}i_1$, is associated with one or more corresponding Unit Expressions denoted as $\langle \mathsf{UE}i_1, \mathsf{UE}i_2, \ldots, \mathsf{UE}_{i_m} \rangle$.

Within our VSR Model $\mathcal{M}$, the input comprises a set of visemes units (UV), which can encompass (i) individual letters and digits, (ii) words, or (iii) complete sentences and phrases. The model generates expression units (UE) as its output. The underlying concept of this model involves establishing a complex m-to-n relationship between visemes and expressions, which may represent phonemes or textual representations. This relationship is characterized by both implicit and explicit mappings in a "many-to-many" manner. Examples of these mappings include $\langle \mathsf{UV}_i \rightarrow \langle \mathsf{UE}_j \rangle$, $\langle \mathsf{UV}_i \oplus \mathsf{UV}_j \rangle \rightarrow \langle \mathsf{UE}_k \rangle$, and $\langle \mathsf{UV}_i \rightarrow \langle \mathsf{UE}_j \oplus \mathsf{UE}_k \rangle$, where $\oplus$ denotes an operator for sequential or parallel composition. Figure 4.5 visually illustrates this concept.

The fundamental aim of VSR is to derive the target sentences, denoted as $\mathsf{Trans}_{\mathcal{V}_{\mathsf{new}}}$, from visual input data comprising videos ($\mathcal{V}$) and models ($\mathcal{M}$). The overarching objective is to develop an efficient solution for VSR, ultimately generating voice or captions corresponding to the pertinent video content. The recognition workflow is structured around three pivotal phases: the "Building Dataset Phase," the "Building Model Phase," and the "Testing Model Phase." Each of these stages plays a critical role in ensuring the systematic progression of the system towards its goal with efficiency and efficacy.

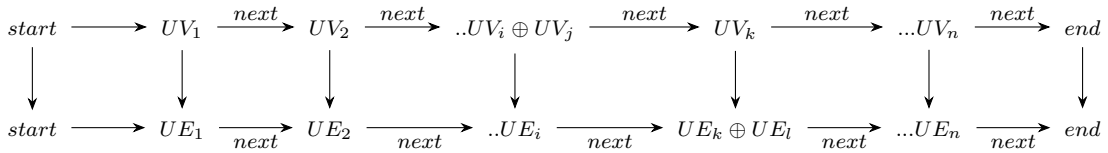


FIGURE 4.5: Synchronize both unit lip movements (UV) and Unit Expression (UE).

4.3.2 Workflow for Constructing Prediction Models

The `BlidAVS10` framework, depicted in Figure 4.2, employs a combination of CNN and ViT models to construct lip movement expressions. This approach aims to delve deeply into nuanced lip movements and extract meaningful features for accurate prediction.

`BlidAVS10` consists of four key components: *Dataset (Preprocessing and Data Augmentation)*, *Mouth detection and tracking*, *Model Recognition*, and *Evaluation Metrics*.

$$\text{BlidAVS10} = \langle \mathcal{D}, P, Aug^{VSR}(D), M_i^{VSR}, m_{i_j} \rangle$$

Where :

- Dataset, denoted as $\boldsymbol{D}$ is organized$(X_{ij};\ Y_i)$, with $i \in [1;\ n]$ and $j \in [1;\ J]$. $\mathcal{D}$ is splited into the training instances $\mathcal{D}_{train}$ and the testing instances $\mathcal{D}_{test}$; each instance is a feature vector. The input P must ensure a higher coverage of $\mathcal{D}$.
- m_{i_j}: a set of measures m_1, ..., m_n expressing the evaluation metrics related to each model selection, e.g. *model accuracy*.
- $\boldsymbol{M_i^{VSR}}$**:** Is the VSR model corresponding to a given metric m_{i_j} produced by CNN e.g. *VGG16 model*, *InceptionV3* or *ViT*.

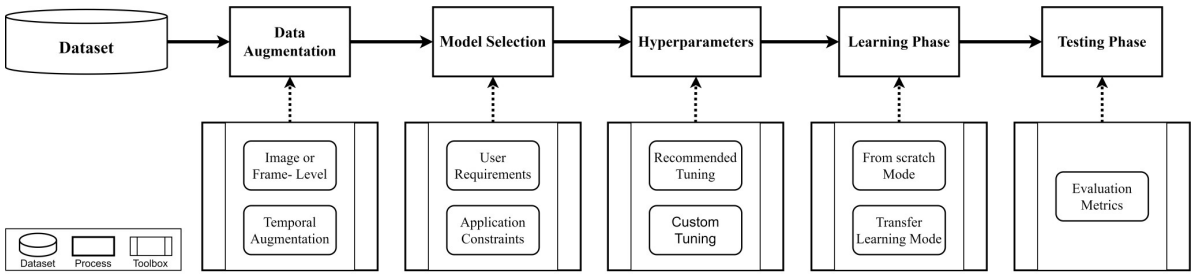


FIGURE 4.6: Building Models Workflow

- P**:** Data preprocessing to ensure optimal performance when used in VSR algorithms; e.g. *Grayscale* and *Normalization*.
- $Aug^{VSR}(D)$**:** Data Augmentation selection; e.g. *Shifting Vertical/Horizontally*, *Correlation Analysis*, *Rotation*.

In this study, we develop a supervised learning problem with arbitrary loss functions. Let Y_i be the prediction of the $\mathcal{M}$ for the specific dataset D_i= $\{V_{i_1},...,V_{i_k}\}$. All these V^k contribute to predict the value of Y_i and minimize the loss $\mathcal{L}(Y, \overline{Y})$ that is calculated with use of *mean-squared error* between the predicted voice or caption Y and real voice or caption $\overline{Y}$. The training objective in a DL model is first the minimization of the desired loss. The objective in this first step is to adjust the model parameters with respect to the real voice or caption value (i.e. true prediction), as seen in Equation 4.1.

$$\underset{M_i^{m_{i_j}}}{Min}\ \mathcal{L}(M_i^{m_{i_j}}(D_i), y) \tag{4.1}$$

In our framework, we posit that the presence of various models facilitates the generation of voice or caption corresponding to the relevant video content. To this end, we propose four models: **(i)** our proposed CNN, **(ii)** VGG16, **(iii)** InceptionV3, and **(iv)** ViT. To quantify the efficacy of our models, one needs only to examine the Residual Sum of Squares (RSS):

$$RSS(k) = 1/k \sum_{t=1}^{k} (Y^{(t)} - \overline{Y}^{(t)})^2 \tag{4.2}$$

In the building models phase, we offer control flow support based on our dataset outlined in the subsequent chapter. Our workflow underscores *learning model*, *testing model*, and *deployment* as service components. Within each phase, the control flow of a process delineates various components available as tools for data scientists and ML researchers. To construct the prediction model, users must adhere to the workflow delineated in Figure 4.6. This workflow comprises a sequence of execution steps.

In summary, all the calls required to analyze the trace are as follows:

- **Step (A):** A data augmentation should be launched to balance a given dataset.
- **Step (B):** one should express the *requirement analysis* by specifying the users' requirements and application constraints to select the appropriate learning models.
- **Step (C):** model hyper-parameters setting must be chosen based on both modes: parameters recommend or customize parameters values.

- **Step (D):** launch the learning model step, if the user needs to use a learning transfer model mode, they shall just load the model and calibrate its parameters.
- **Step (E):** the model testing according to the evaluation metrics selected by users (i.e. precision, accuracy). Since each evaluation metric is not achieved, the user can select other models or perform model settings.

The architecture of our framework can be modified to visualize its structures, signatures, and components, allowing manual adaptation by developers. The UML Component diagram, illustrated in Figure 4.7, provides a logical view of the key functional components of BlidAVS10. While most of these components are also found in DL libraries like TensorFlow and Keras, we have adapted and extended them. According to the legend of the diagram, unchanged components are depicted in blue, while adapted/extended components are shown in yellow (cf. Service 4 in Figure 4.2).

Our framework's architecture can be utilized in a declarative manner, where users simply utilize it as is to produce the desired voice or caption corresponding to the relevant video (cf. Service 3 in Figure 4.2). Alternatively, designers can opt for an explicit customization approach, wherein they can modify the architecture of the CNN Learning Algorithm.

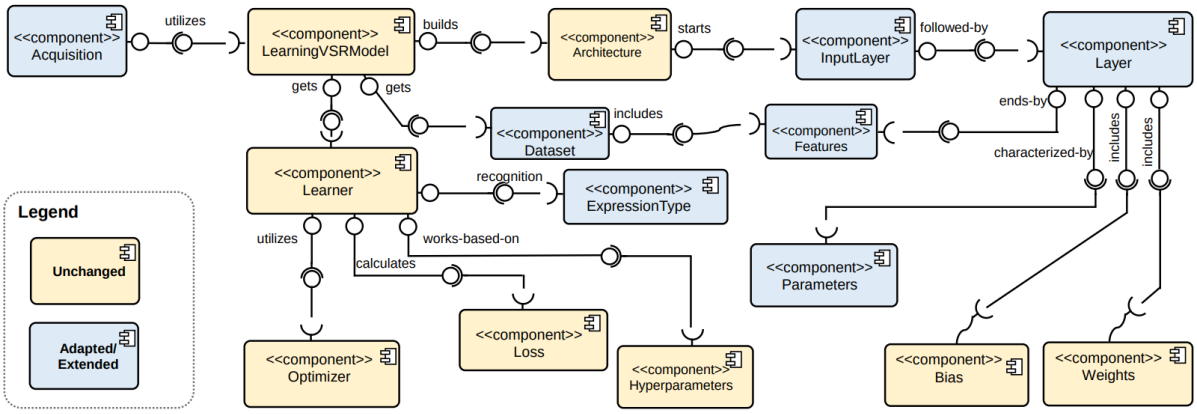


FIGURE 4.7: The component diagram illustrating the logical architecture view of the proposed framework.

4.4 BlidaAVS10 Framework Services

In this section, we delve into the core functionalities of the BlidaAVS10 framework, unravelling the its comprehensive services. From the inception of video data acquisition to the deployment of Deep Learning (DL) models, each service plays a pivotal role in the seamless progression of Visual Speech Recognition (VSR). The following elucidation offers a detailed exploration of the framework's services, shedding light on their individual contributions and the collaborative synergy that underpins the entire process.

4.4.1 Service 1: Dataset Creation and Analysis

Service 1, the cornerstone of our framework, is dedicated to the creation and analysis of datasets expressed as a set of videos. This service revolves around the meticulous

acquisition process of video input, strategically designed to capture a diverse range of visual speech patterns. The resulting dataset forms the foundation for subsequent stages in the BlidaAVS10 framework.

`Video Acquisition Process:` Service 1's video acquisition process is meticulously crafted to capture a representative sample of visual speech instances. By focusing on individual phonemes or visemes, our dataset aims for granularity, enabling nuanced analysis for enhanced model training.

`Dataset Structuring:` This service extends beyond mere video collection. It facilitates the organization of acquired instances into a structured dataset. Instances are stored under a unified framework, promoting seamless reuse, comparison, and storage. Users benefit from a standardized structure that eases dataset management, fostering efficient downloading and uploading processes.

`Interactivity and Accessibility:` Service 1 enhances interactivity by allowing users to engage with the dataset creation process. The user-friendly interface supports intuitive interactions, enabling researchers to actively participate in the formation of a comprehensive dataset. Accessibility is emphasized, ensuring that the dataset aligns with specific research needs and objectives.

4.4.2 Service 2: Mouth Region Detection and Tracking

Service 2 focuses on empowering developers with tools for preprocessing video input, precise mouth region detection, and frame aggregation. This service, illustrated in Figure 4.2, plays a pivotal role in preparing data for subsequent stages.

`Preprocessing Input Data:` Developers leverage Service 2 to preprocess raw video input. This step involves essential tasks such as format standardization and quality enhancement. The preprocessing phase sets the stage for accurate and reliable mouth region extraction.

`Mouth Region Extraction:` Service 2 employs the renowned dlib library for efficient mouth region detection and tracking. Leveraging advanced facial landmark detection capabilities, dlib ensures the accurate extraction of the mouth region within each frame. This step is crucial for isolating visual speech cues.

`Frame Aggregation:` After extracting mouth regions, Service 2 aggregates frames, determining their count for further analysis. This aggregation process simplifies subsequent stages by providing a concise representation of visual speech patterns. Developers benefit from a streamlined dataset for model customization.

4.4.3 Service 3: VSR Model Selection and Customization

Service 3 caters to developers, offering a versatile platform for selecting and customizing Visual Speech Recognition (VSR) models. This service empowers researchers to tailor models according to specific use cases, with a focus on attributes, CNN architecture, and hyper-parameter customization.

`Attribute Customization:` Developers can fine-tune attributes crucial for VSR models. This includes adjusting attributes related to the visual representation of lip movements, enabling a personalized approach that aligns with the nuances of different research scenarios.

`CNN Architecture Customization:` Service 3 provides developers with the flexibility to customize Convolutional Neural Network (CNN) architectures. This customization

extends beyond conventional models, allowing researchers to experiment with architectures optimized for their unique datasets and objectives.

`Hyper-parameter Customization:` Fine-tuning hyper-parameters is a critical aspect of model optimization. Service 3 facilitates the customization of hyper-parameters, ensuring that VSR models align with specific performance criteria and user-defined constraints.

4.4.4 Service 4: DL Model Implementation and Evaluation

The final stage of the BlidaAVS10 framework, Service 4, encompasses the implementation, evaluation, and deployment of Deep Learning (DL) models. This service acts as the culmination of the framework, ensuring that meticulously customized models are ready for real-world applications.

`Model Implementation:` Service 4 brings the customized VSR models to life through comprehensive implementation. Leveraging DL libraries such as TensorFlow and Keras, this stage transforms theoretical configurations into functional models, ready for deployment.

`Rigorous Model Evaluation:` Before deployment, Service 4 rigorously evaluates the performance of implemented models. Evaluation metrics, including precision, accuracy, and other relevant benchmarks, provide insights into the model's effectiveness. This iterative process ensures that only robust models progress to the deployment phase.

`Deploying Models for Real-World Applications:` The deployment phase marks the transition from theoretical frameworks to practical utility. Service 4 facilitates the seamless deployment of optimized VSR models, making them accessible for real-world applications. This includes scenarios demanding real-time processing, ensuring the adaptability of the models to diverse user requirements.

In summary, the BlidaAVS10 framework's four services collectively form a dynamic and cohesive ecosystem, empowering researchers to navigate the complexities of Visual Speech Recognition. From dataset creation to model deployment, each service plays a vital role, contributing to the framework's effectiveness in addressing the challenges of this evolving field.

4.5 Conclusion

In conclusion, this chapter has delved into the BlidAVS10 framework, offering a comprehensive exploration of its innovative approach to Visual Speech Recognition (VSR). We have elucidated the four pivotal services provided by BlidAVS10, encompassing the creation of datasets, precise mouth region detection and tracking, flexible customization of VSR models, and the robust building, training, and evaluation of Deep Learning (DL) models.

The theoretical foundations presented in the "Mathematical Theoretical" section have further enriched our understanding, providing a mathematical lens through which we scrutinize the core concepts of dataset creation and the workflow for constructing prediction models within the BlidAVS10 framework.

As we conclude this chapter, it is paramount to acknowledge the accomplishments and advancements showcased by BlidAVS10. However, we also recognize the persistent challenges in the field, including real-time applications, model accuracy, and the dynamic nature of visual speech. These challenges underscore the continuous need for refinement and adaptation in response to the evolving landscape of VSR.

Looking ahead, this conclusion serves as a bridge to future research endeavors. The insights gleaned from BlidAVS10 open avenues for further exploration, ranging from enhanced dataset creation techniques to the optimization of DL models for real-world applications. The iterative nature of research and the dynamic demands of technology necessitate an ongoing commitment to innovation and improvement in the realm of Visual Speech Recognition.

In essence, the "Conclusion" section not only summarizes the achievements and contributions of this chapter but also emphasizes the forward-looking nature of research in VSR. As we move forward, armed with the knowledge and insights gathered here, the exploration into the evolving landscape of Visual Speech Recognition continues, inviting researchers and practitioners to shape the future trajectory of this dynamic field.

Chapter 5

Building the Arabic Audio-Visual Dataset (BlidAVS10)

5.1 Introduction

The endeavor to advance Visual Speech Recognition (VSR) systems, particularly in the context of the Arabic language, necessitates a robust foundation of high-quality datasets. In response to this need, this chapter introduces the meticulous process of constructing the Arabic Audio-Visual Dataset, referred to as BlidAVS10. The significance of a well-crafted dataset cannot be overstated, as it serves as the cornerstone for training, testing, and refining VSR models. BlidAVS10 is designed with a nuanced approach, encompassing the complexities of the Arabic language, which is not only one of the oldest languages globally but also boasts a vast number of speakers.

This chapter embarks on a comprehensive exploration of the methodology and considerations involved in creating BlidAVS10. It delves into the complexities of capturing audio-visual content that is representative of the diverse linguistic landscape within the Arabic-speaking community. The unique phonetic and visemic characteristics of the Arabic language pose specific challenges and opportunities in the dataset creation process, and this chapter aims to shed light on the strategies employed to address these nuances effectively.

As we venture into the details of BlidAVS10, the chapter provides a roadmap for the creation of a dataset that not only captures linguistic diversity but also caters to the visual nuances of speech. From the acquisition of video inputs to the meticulous annotation of phonemes and visemes, each step in the dataset creation process is meticulously examined. The goal is to lay the groundwork for a dataset that not only facilitates the training and evaluation of VSR models but also serves as a valuable resource for the broader research community interested in Arabic linguistics and audio-visual data analysis.

Furthermore, the introduction outlines the organizational structure of the chapter, providing a glimpse into the subsequent sections. These sections will unfold the methodology, challenges, and key considerations in building the Arabic Audio-Visual Dataset, contributing to the advancement of research and technology in the realm of Visual Speech Recognition, with a specific focus on the Arabic language.

5.2 Constructing the Dataset

Commencing the establishment of an Arabic speech recognition dataset ($\mathcal{D}$) posed notable challenges (cf. Service 1 Figure 4.2). Figure 5.1 delineates the dataset construction process, encompassing the following steps:

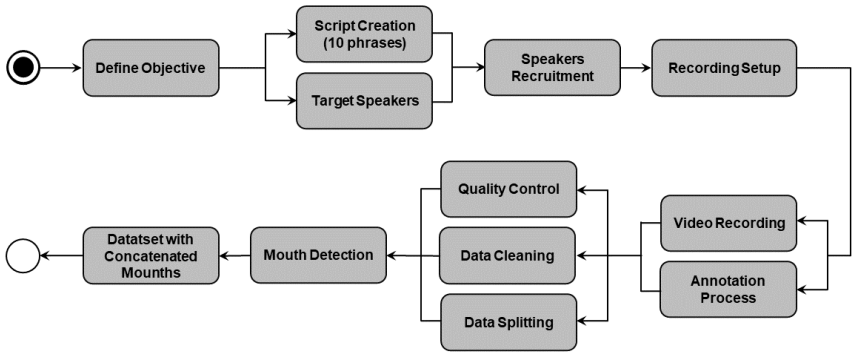


FIGURE 5.1: Building dataset process.

1. ***Define Objective:*** Clearly articulate the overarching research goal of constructing the dataset, emphasizing its pivotal role in advancing Visual Speech Recognition (VSR) research. This objective serves as the foundation for subsequent decisions and actions in the dataset creation process, ensuring alignment with research priorities.

2. ***Select Target Speakers:*** Rigorously identify and choose speakers based on predetermined criteria to ensure the dataset's diversity and representativeness. Consider factors such as age, gender, regional accents, and linguistic proficiency to capture a broad spectrum of variations inherent in natural speech.

3. ***Script Generation:*** Formulate a comprehensive collection of standardized scripts and phrases incorporating diverse linguistic elements, guaranteeing representation of phonetic and visemic diversity. These scripts establish the linguistic foundation for recording sessions, ensuring the creation of a well-rounded dataset.

4. ***Participant Recruitment:*** Execute a meticulous participant recruitment process, adhering to the predefined criteria for dataset contributors. Strive for diversity among participants, reflecting the natural variability present in the target speech community.

5. ***Recording Setup:*** Create a controlled recording environment by optimizing variables such as lighting conditions, camera angles, and audio quality. A well-prepared recording setup ensures high-quality data capture, minimizing external influences on the recorded content.

6. ***Video Recording:*** Conduct recording sessions with participants, capturing facial expressions, lip movements, and audio simultaneously. Ensure that the recorded data comprehensively represents the linguistic and visual nuances intended for analysis in VSR.

7. ***Annotation Process:*** Implement a meticulous annotation process to label and tag spoken words or phrases, aligning them precisely with corresponding mouth movements. Accurate annotation is crucial for training and evaluating VSR models effectively.

8. ***Quality Control:*** Enforce rigorous quality control measures to identify and rectify any issues in the recorded data. Consistency and accuracy are paramount, and any discrepancies or anomalies should be addressed during this phase.

9. ***Data Cleaning:*** Thoroughly clean and preprocess the dataset to eliminate any noise, artifacts, or inconsistencies. A clean dataset forms the basis for reliable VSR model training and evaluation.

10. ***Data Splitting:*** Partition the dataset into suitable subsets designated for training, validation, and testing. Ensure equitable speaker distribution across subsets to preserve diversity and bolster the generalizability of trained models.

11. ***Mouth Detection:*** Utilize advanced computer vision methodologies to detect and isolate mouth regions within individual video frames. This step is crucial for isolating pertinent visual cues pivotal for VSR analysis.

12. ***Creation of Concatenated Mouth Datasets:*** Generate finalized datasets featuring concatenated mouth movements, priming them for integration into the training and evaluation workflows of VSR models. This amalgamated dataset format streamlines the incorporation of visual speech cues into model learning processes.

Below, we offer an exhaustive breakdown of the most important of these steps, delving into specific considerations and methodologies to ensure a comprehensive understanding of the dataset creation process.

5.3 Data Collection Setup

Collecting top-notch audio-visual data for lipreading requires meticulous planning. We carefully selected suitable speakers, prepared standardized scripts for reading aloud, and recruited a diverse group of participants. In a controlled recording environment, we optimized factors like lighting, camera angles, and audio quality. During recording, we captured facial expressions, lip movements, and audio simultaneously. The annotated data underwent rigorous quality checks and cleaning processes to ensure accuracy.

5.4 Recording Environment

Creating a controlled yet diverse recording environment is crucial for collecting representative data. The following considerations were taken into account:

- **Lighting Conditions:** Adequate lighting, employing [type of lighting], was meticulously arranged to facilitate clear capture of facial expressions and lip movements. This precautionary measure aimed to prevent harsh shadows or overexposure, ensuring optimal visibility of the speaker's features.

- **Background:** A neutral background devoid of distracting elements was chosen to focus on the speaker's facial features. This helps in maintaining the viewer's attention on lip movements.

- **Recording Room:** The recording sessions were conducted in a controlled environment, such as a quiet and soundproof recording room [or other specific environment details]. This choice served the dual purpose of minimizing ambient noise, crucial for clean audio signals, and providing a consistent setting for capturing visual data.

5.4.1 Equipment Used

The data collection setup utilized common equipment to ensure accessibility and flexibility. The following were employed:

- **Smartphones:** The camera feature of smartphones served as the primary recording device, offering portability and ease of use.
- **Tripods and Mounts:** Sturdy tripods and mounts were strategically used to stabilize both the smartphone camera and any additional microphones, minimizing unwanted movements during recording.

5.4.2 Challenges and Resolutions

Despite meticulous planning, additional challenges arose during the data collection setup. Each challenge posed unique hurdles that demanded prompt and effective resolutions. Here, we outline these challenges along with the corresponding strategies employed to ensure a smooth and successful data collection process:

- **Audio Interference:** Occasional interference or noise in the audio signal was observed. This was effectively mitigated by implementing measures such as using shielded cables or adjusting microphone placement. Regular checks were conducted to ensure optimal audio quality throughout the recording sessions.
- **Participant Comfort:** Ensuring participant comfort during recording sessions was paramount. Breaks were strategically incorporated into the schedule, and clear communication channels were maintained to address any concerns or discomfort expressed by the participants. This proactive approach not only enhanced the overall experience for the participants but also contributed to the quality of the captured data.
- **Technical Glitches:** Unforeseen technical glitches are an inevitable aspect of any setup. To address this, a well-defined protocol for handling technical issues was established. A backup system was also in place to prevent data loss, ensuring that the data collection process remained robust and reliable.
- **Participant Consent:** Convincing female participants to consent to recording, especially considering that the datasets would be in the form of videos shareable with researchers, presented a unique challenge. Strategies were devised to address this concern and ensure ethical and inclusive participation.

The careful consideration of equipment, environment, and proactive resolution of challenges contributed to the successful setup of the data collection process for the audio-visual dataset. This dedication to overcoming challenges underscores the commitment to obtaining high-quality and reliable data for furthering research in Visual Speech Recognition.

5.4.3 Video Recording Process

The process of constructing the Arabic dataset involved meticulous planning and execution to ensure the quality and diversity essential for effective visual speech recognition (VSR). We carefully selected 10 sentences, commonly used in everyday Arabic communication, aiming for a broad vocabulary and expressive range. Table 5.1 provides a comprehensive list of these sentences, their English pronunciations using the Latin alphabet, and corresponding English translations.

TABLE 5.1: Ten daily phrases selected.

The initial column presents the Arabic sentences, while *the second column* showcases their English pronunciation represented in the Latin alphabet. *The third column* offers the corresponding English translations.

Phrase in Arabic	Phrase Syllables	Phrase in English
السلام عليكم	Assalamu alaikum	Hello
صباح الخير	Sabah al-khair	Good morning
متأسف جدا	Mataasifun jiddan	I am so sorry
ماذا تفعل	Maadha taf'al	What do you do
هل انت بخير	Hal anta bi-khair	Are you fine
تعال معي	Ta'al ma'i	Come with me
كيف حالك	Kayf halak	How are you
كن شجاعا	Kun shuja'an	Be brave
أريد الدخول	'Ureed al-dkhul	I want to come in
شكرا لك	Shukran lak	Thank you

In this project, our focus was on providing phrases rather than individual words, a deliberate choice aimed at facilitating the training process for automatic Visual Speech Recognition (VSR) systems. To ensure the dataset's richness and variability, each selected speaker was given specific instructions to repeat each phrase at least 10 times. This repetition was intended to capture a wide range of articulations and accentuate the natural variability in pronunciation that occurs across different speakers.

For the video recordings, we opted to use smartphones due to their widespread availability and ease of use. The recording setup involved placing the smartphone approximately 1 meter away from the speaker to capture clear and consistent audio and visual data. To create a natural and dynamic recording environment, the recording room was set up with an open balcony door to leverage natural lighting. While the background color was not standardized across all recordings, intentional variation was introduced to mimic real-world scenarios where lighting conditions may vary.

To maintain consistency in the dataset, we employed a single camera angle throughout the recordings, with the speaker facing directly toward the camera. This decision was made to simplify subsequent data processing and analysis, ensuring that all recorded videos followed a uniform format. By maintaining a consistent camera angle, we aimed to minimize potential confounding factors and maintain dataset integrity for accurate analysis and model training.

5.5 Video Details

In our pursuit of high-quality recordings, meticulous attention was given to the video capture settings. The camera was precisely configured to record videos at a resolution of 1920x1080 pixels, coupled with a frame rate of 30 frames per second. This carefully chosen setup guarantees that the resulting videos not only possess a crisp and detailed resolution but also exhibit smooth and precise motion. Such video details are paramount for accurate

visual speech recognition, ensuring that subtle lip movements and facial expressions are faithfully captured and analyzed.

In our endeavor to construct a comprehensive dataset for Arabic visual speech recognition, we took careful steps to ensure diversity among participants. This involved selecting individuals with varying characteristics such as age, gender, and facial hair.

Participants were chosen across a wide age range, spanning from 8 to 45 years old, to capture the diversity in speech patterns across different age groups. Both male and female participants were included, with six males and three females, ensuring gender representation.

To further enrich the dataset, we included individuals with different facial hair styles, considering the potential impact of facial hair on lipreading accuracy. This selection encompassed participants with and without beards and mustaches, enhancing the dataset's relevance to real-world scenarios.

Figure 5.2 provides a visual depiction of some participant images, highlighting the dataset's diversity in age, gender, and facial hair styles. These images reflect our commitment to inclusivity and showcase the varied representation within the dataset.

FIGURE 5.2: Images of some participants demonstrating differences in age, gender, the presence of beards and mustaches, and variations in lighting and background.

These visual representations serve to underscore our dedication to creating a dataset that authentically reflects the diversity encountered in real-world scenarios. The inclusion of participants with different characteristics enriches the dataset, making it more robust and applicable to a wide range of visual speech recognition scenarios. The subsequent sections delve into the data processing methods and the specific neural network architectures employed in this study.

5.6 Video Processing

Upon completion of the video collection process, a crucial step involved filtering out unnecessary clips, excluding periods that did not contain frames of the spoken sentences as per the specified conditions. It was observed that the correct videos did not exceed two seconds in duration. To streamline this process, a custom program was developed utilizing the `ffmpeg` tool [**ffmpeg**] for the cutting process. The program utilized a file containing the starting time for each video and saved the results in the ".mp4" format. Each file was labelled according to a specific convention:

The file name format consisted of the following elements:

- The first letter (**s**) representing the speaker.
- The speaker's number.
- A letter (**p**) representing the sentence.
- The sentence number.
- The first letter of the utterance (**u**) followed by its number.

As an example, the file name for the first speaker pronouncing the first sentence and first utterance would be `S01p01u01.mp4`.

A summary of the specifications for the proposed `BlidAVS10` dataset is provided in Table 5.2.

Table 5.2: Summary of the proposed `BlidAVS10` dataset specifications.

Category	Specifications
Language	Arabic Language Modern Standard (MSA)
Task	Sentences
Nmb. of Utterances	10
Nmb. of Participants	9
Gender	6 Males and 3 Females
Nmb. of Utterance Repeat	More than 10
Face View	Frontal
Multi-Speakers	No
Camera Type	Smartphone Camera
Acquired Data	Video (MP4 format)
Resolution	Full HD (1920×1080 Pixels)
Frame Rate	30 fps
Environment	Indoor illumination
Background	Simple

This dataset, labeled as `BlidAVS10`, is meticulously designed to meet specific criteria, ensuring its relevance for visual speech recognition tasks. The specifications encompass language, task details, participant information, and technical aspects of video acquisition.

5.7 Available Dataset Types for Research

In the pursuit of advancing Visual Speech Recognition (VSR) research, we provide three meticulously curated dataset types, each tailored to address specific research requirements. These datasets not only contribute valuable insights to the field but also serve as essential resources for fostering collaboration and innovation in the scientific community.

1. **Raw Dataset:** This foundational dataset, comprising 1383 videos meticulously organized into folders, captures participants expressing the same sentence in diverse contexts. The richness of this collection not only serves as a cornerstone for advancing Visual Speech Recognition (VSR) but also holds immense potential for a myriad of research avenues beyond VSR. By sharing the complete raw dataset, we aim to empower researchers to explore applications such as face detection, voice and audio recognition, or any domain where visual speech data can be harnessed. This inclusive approach encourages versatility in research applications and invites innovative perspectives that extend beyond the immediate scope of VSR, Table 5.3 shows the frames and videos for each phrase in our datasets provided.

2. **Processed Dataset:** Tailored for researchers employing models such as LSTM, GRU, RNN, etc., this dataset is designed to streamline the research process by providing prepared and processed data. Comprising 10 folders, each corresponding to a sentence, the dataset offers meticulously cropped mouth images. With 60 images generated from each video, the total collection amounts to 82,980 images (1383×60). This prepared dataset aims to expedite the task of dataset preparation for lipreading research. By offering pre-processed data, we seek to minimize the time and effort researchers typically invest in data preparation, enabling them to focus more on the core aspects of advanced model training and analysis of mouth movements.

3. **Ready-for-Input Dataset:** Streamlining the integration of models such as VGG16, InceptionV3, ResNet, etc., this dataset consists of 10 folders, each representing one of the sentences. Annotations for 49 frames of a single video are consolidated into one image, offering a ready-to-use input for models and encouraging efficient experimentation with various visual speech recognition architectures.

Sharing these datasets is paramount to fostering collaborative research and enabling the scientific community to collectively push the boundaries of Visual Speech Recognition.

Phrase	**Frames**	**Number of Items**
p1	7200	120
p2	7560	126
p3	9000	150
p4	8520	142
p5	8820	147
p6	9120	152
p7	8160	136
p8	7800	130
p9	8400	140
p10	8400	140

TABLE 5.3: Number of frames in the processed dataset and items (videos/concatenated images) in the other 2 datasets for each phrase

5.8 Conclusion

The meticulous planning and execution of the data collection setup, encompassing equipment selection, environmental considerations, and proactive resolution of challenges, have laid the foundation for a robust audio-visual dataset. By addressing technical nuances and participant comfort, we have ensured the acquisition of high-quality data essential for advancing research in Visual Speech Recognition. The commitment to overcoming challenges underscores our dedication to maintaining the integrity and reliability of the dataset, paving the way for meaningful contributions to the field.

Chapter 6

Features Extraction: Mouth Detection and Tracking

6.1 Introduction

Visual Speech Recognition (VSR) stands at the forefront of cutting-edge technology, aiming to bridge the gap between visual input and linguistic understanding. At the heart of this sophisticated process lies the pivotal role of feature extraction, a cornerstone that shapes the precision and efficacy of VSR models. This chapter embarks on an in-depth exploration of the nuanced world of feature extraction, honing in on the specialized domain of Mouth Detection and Tracking.

The significance of feature extraction in VSR cannot be overstated; it serves as the conduit through which raw visual cues are transformed into meaningful information. The efficacy of model training and the subsequent accuracy of predictions hinge on the aptness of feature extraction methods employed. As we navigate through the complexities of this crucial component, our primary focus will be on the processes involved in detecting and tracking the mouth—a key region that encapsulates the subtleties of speech.

Mouth Detection and Tracking, a subset of feature extraction, emerge as central themes in this exploration. These processes play a pivotal role in deciphering the patterns of visual speech, where nuanced lip movements encode linguistic nuances. This chapter unfolds layers of understanding, detailing the methodologies, tools, and considerations that underpin effective feature extraction in the realm of VSR.

As we unravel the detailed of Mouth Detection and Tracking, we delve into the practical applications of tools like Dlib and the implementation of state-of-the-art models like YOLO for real-time detection. The chapter extends beyond theoretical discussions, offering practical insights into the training processes, comparisons between different approaches, and the subsequent steps involved in data processing.

By delving into the rich tapestry of feature extraction, this chapter endeavors to empower researchers, practitioners, and enthusiasts in the field of VSR. The specialized domain of Mouth Detection and Tracking aims to equip you with the knowledge and tools necessary to push the boundaries of visual speech understanding.

6.2 Importance of Feature Extraction in VSR

Visual Speech Recognition (VSR) is a multidimensional process that involves deciphering spoken language through visual cues. At its core, feature extraction acts as the linchpin, transforming raw visual data into meaningful information that VSR models can interpret.

The importance of this intermediary step cannot be overstated, particularly in the nuanced domain of lipreading.

Lipreading, a subset of VSR, relies on extracting features from the visual input, primarily focusing on the movements of the mouth. The human mouth, with its subtle articulations and gestures, becomes a canvas that encodes linguistic nuances. Effective feature extraction is crucial for unveiling these subtleties, providing the VSR model with the necessary insights to accurately decode spoken language.

The significance of feature extraction in VSR lies in its ability to distill complex visual information into manageable and informative elements. By highlighting key aspects of the visual input, feature extraction simplifies the learning process for VSR models, enabling them to recognize patterns, phonemes, and linguistic nuances. This transformative process elevates the accuracy and efficiency of VSR models, making them adept at understanding the visual speech.

In the realm of lipreading, where traditional audio cues may be insufficient or absent, the visual modality gains prominence. Feature extraction becomes the mechanism through which visual cues, especially those from the mouth region, are translated into actionable data. It acts as a crucial preprocessing step, paving the way for effective model training and subsequent accurate predictions.

As researchers and practitioners strive to enhance the capabilities of VSR systems, the quest for optimal feature extraction methodologies becomes paramount. The adaptability of feature extraction techniques to diverse linguistic contexts, variations in articulation, and environmental conditions is a testament to its pivotal role in the success of VSR applications.

This chapter navigates the multifaceted landscape of feature extraction in VSR, with a keen focus on the specialized domain of Mouth Detection and Tracking. Our exploration extends beyond theoretical discussions, providing practical insights into the methodologies and tools employed to unravel the complexities of visual speech. By understanding the importance of feature extraction, we lay the foundation for advancements in the field, empowering VSR systems to decode spoken language with unprecedented accuracy and sophistication.

6.3 Region of Interest (ROI) and Dlib

In the realm of visual speech processing, accurately pinpointing the Region of Interest (ROI) is pivotal for effective feature extraction, particularly when focusing on the nuances of mouth movements. Dlib, a powerful computer vision library, emerges as a cornerstone in this process, offering robust tools for precise and efficient ROI detection.

The concept of ROI is fundamental, as it defines the specific area within an image or frame where critical visual information, in this case, the mouth region, is concentrated. Dlib's sophisticated algorithms and pre-trained models excel in identifying and isolating the ROI with remarkable accuracy.

The process of ROI detection using Dlib involves multiple stages. Initially, facial landmarks are identified, providing key points that act as anchors for subsequent analyses. The mouth region, being one of the primary points of interest, is delineated based on these facial landmarks.

Dlib's versatility extends to handling variations in pose, lighting conditions, and facial expressions, making it a robust choice for ROI detection in diverse scenarios. Our approach

simplifies mouth tracking by ensuring that each frame contains a single mouth without occlusions, thus eliminating the complexity associated with multiple mouth detection or occluded views.

The chapter offers a granular exploration of the step-by-step procedures involved in leveraging Dlib for ROI identification. Practical insights into parameter tuning, model selection, and potential challenges are discussed to empower researchers and practitioners in implementing robust ROI detection methodologies.

Emphasizing Dlib's practical applications, the chapter not only delves into the theoretical aspects of ROI detection but also provides hands-on guidance, ensuring a comprehensive understanding of its strengths and limitations. The utilization of Dlib for ROI detection serves as a foundational step in the subsequent feature extraction processes, laying the groundwork for precise and context-aware visual speech recognition.

6.4 Training YOLO for Real-Time Mouth Detection

This section outlines the comprehensive process of training a YOLO (You Only Look Once) model tailored for real-time mouth detection. The training phase encompasses vital aspects such as dataset preparation, model architecture, hyperparameter tuning, and evaluation metrics. Practical considerations for diverse scenarios, lighting conditions, and facial expressions are highlighted to ensure the model's robustness in real-world applications.

6.4.1 Dataset Preparation

The foundation of YOLO training lies in meticulous data preparation. The chapter initiates by detailing the creation of a dedicated dataset for YOLO training. Leveraging the dataset introduced in the previous chapter, 125 frames are randomly selected from different videos. These frames are then split into train and test sets, allocating 80% for training and 20% for testing. To enrich the dataset, annotation is performed using the website tool makesens.ai. The process ensures a diverse and representative dataset, setting the stage for effective YOLO training.

6.4.2 Architecture of YOLOv5

The architecture of YOLOv5 is presented with an in-depth overview, accompanied by necessary information and illustrative figures. The model's underlying principles enabling real-time mouth detection are elucidated. Key hyperparameters are dissected, providing insights into their impact on performance. The chapter offers practical guidelines for parameter tuning, empowering researchers to adapt the YOLO model to specific Visual Speech Recognition (VSR) requirements.

6.4.3 Results

The section includes a comprehensive presentation of the results obtained through the YOLO training process. Performance metrics, including accuracy, precision, and recall, are analyzed to provide a nuanced understanding of the model's capabilities. Visualizations, such as detection outputs on sample frames, enhance the interpretability of the results.

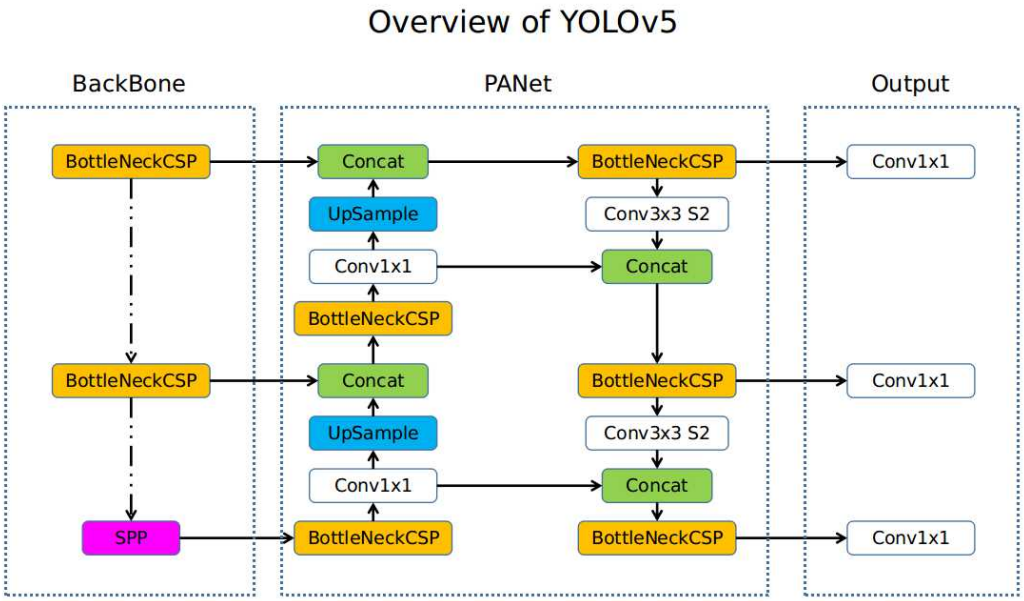


FIGURE 6.1: YOLOv5 architecture.

6.4.4 Discussion

The chapter concludes with a discussion that guides researchers in interpreting and contextualizing the results. Considerations for real-world scenarios, including speed-accuracy trade-offs, are addressed. The discussion serves as a bridge between the theoretical understanding of YOLO and its practical implementation, offering insights for optimizing the model based on specific VSR use cases.

This section serves as a practical guide, providing researchers and practitioners with a step-by-step approach to integrating real-time mouth detection into their VSR systems using YOLO. The meticulous attention to dataset preparation, model architecture, and evaluation metrics ensures a holistic understanding of the training process and its implications in real-world applications.

This section functions as a practical guide, bridging the theoretical understanding of YOLO with hands-on implementation details. Researchers and practitioners aiming to integrate real-time mouth detection into their VSR systems will find valuable insights and actionable steps for leveraging YOLO effectively.

6.5 Comparison between Dlib and YOLO Approaches

This section conducts a meticulous comparative analysis between the Dlib and YOLO approaches, offering insights into their respective strengths and limitations in the context of mouth detection for Visual Speech Recognition (VSR).

6.5.1 Algorithm Overview

The video processing algorithm utilizes both YOLOv5 and Dlib for mouth detection. The key steps include:

Algorithm Overview

Initialize YOLOv5 Model:

- ***Load Model:*** Load YOLOv5 model from specified path.

Define Utility Functions:

- ***shape_to_list:*** Convert dlib shape to a list of coordinates.
- ***process_with_yolo:*** Process frame using YOLOv5, extract relevant regions.
- ***process_with_dlib:*** Process frame using Dlib, extract lip regions.

Main Loop Over Video Folders:

- ***Iterate Folders:*** Iterate through video folders ('p1', 'p2', 'p3').

Process Each Video:

- ***Open Video:*** Open video, initialize frame buffers, and processing time counters.

Frame Processing Loop:

- ***Read Frames:*** Read frames until no frames are left.
- ***Measure Time:*** Measure processing time for YOLOv5 and Dlib for each frame.
- ***Accumulate Times:*** Accumulate total processing times.

Print Processing Times:

- ***Print:*** Print total processing times for YOLOv5 and Dlib for each video.

Release Video Resources:

- ***Release:*** Release resources associated with the processed video.

FIGURE 6.2: Algorithm for Video Processing using YOLOv5 and Dlib to detect and crop the mouth region.

6.5.2 Processing Time Results

The processing times for YOLOv5 and Dlib on different videos are summarized in Table 6.1.

Video	**YOLOv5 Processing Time (seconds)**	**Dlib Processing Time (seconds)**
s7p01u01.mp4	16.62	54.45
s7p01u02.mp4	16.64	55.27
s7p02u01.mp4	17.03	55.74
s7p03u02.mp4	16.32	54.09
Average	16.65	54.64

TABLE 6.1: Processing Times for YOLOv5 and Dlib on Different Videos

The average processing time for YOLOv5 is 16.65 seconds, while for Dlib, it is 54.64 seconds. These results indicate a substantial difference in processing speed between the two approaches. YOLOv5 demonstrates a notable advantage in terms of faster computation, being approximately three times quicker than Dlib on average for the given set of videos.

6.5.3 Discussion

The processing time results indicate that YOLOv5 exhibits superior computational efficiency in the context of mouth detection for Visual Speech Recognition (VSR), delivering faster processing times compared to Dlib. This efficiency is particularly noteworthy for applications that prioritize real-time analysis and responsiveness.

The efficiency of YOLOv5 in processing frames suggests its suitability for real-time applications or scenarios where rapid analysis of visual data is crucial. Conversely, the longer processing times observed with Dlib may be a consideration in situations where real-time processing is not a strict requirement and precision in landmark detection is prioritized over speed. This highlights the importance of carefully weighing the trade-offs between speed and precision based on the specific needs and constraints of the target VSR application.

While the efficiency of YOLOv5 is a compelling factor, a comprehensive assessment should consider additional aspects beyond processing speed. Accuracy and resource utilization are crucial dimensions to weigh in the decision-making process.

6.5.3.1 Accuracy Consideration

It is imperative to conduct a thorough evaluation of the accuracy of mouth detection provided by both YOLOv5 and Dlib. A comparative analysis of the precision and recall rates can offer insights into the trade-offs between speed and accuracy. This assessment is vital, especially in VSR applications where reliable and precise mouth detection is essential for accurate speech recognition.

6.5.3.2 Resource Utilization

Resource utilization, including memory usage and hardware requirements, should be considered alongside processing speed and accuracy. Different applications may have varying constraints and preferences regarding resource consumption. Understanding the resource utilization patterns of each approach contributes to making an informed decision based on the specific requirements of the target VSR application.

6.5.3.3 Trade-offs Between Speed and Precision

Exploring the trade-offs between processing speed and precision is crucial in determining the most suitable approach. Depending on the application's objectives and constraints, a balanced approach that optimally meets the speed and accuracy requirements may be sought. Identifying the acceptable trade-offs ensures that the chosen method aligns with the application's goals.

In conclusion, while YOLOv5 showcases notable advantages in terms of processing speed, a comprehensive decision should be informed by a detailed analysis of accuracy, resource utilization, and the trade-offs between speed and precision. Further research and experimentation are recommended to fine-tune the selection of the most suitable approach for specific Visual Speech Recognition applications. .

6.6 Data Processing and Concatenation

This section provides a detailed exploration of the post-detection processing steps applied to the data, emphasizing normalization, cleaning, and data augmentation techniques. A pivotal aspect of preparing the dataset for model training is the incorporation of concatenation, particularly focusing on the arrangement of cropped mouth regions. This process is designed to enrich the contextual information available to the model, enhancing its robustness and generalization capabilities.

6.6.1 Concatenation

To convert temporal data into a spatial format, the sequence of mouth frames undergoes a meticulous arrangement. This sequential organization, exemplified in Figure 6.3, holds significant importance for methodologies such as LSTM and GRU, where the temporal sequence of data plays a crucial role. Building upon this fundamental concept, our innovative approach involves the concatenation of these ordered frames into a single image, strategically designed with dimensions of 224 x 224 pixels. Notably, this transformation intentionally excludes the remaining 11 frames to streamline the data representation.

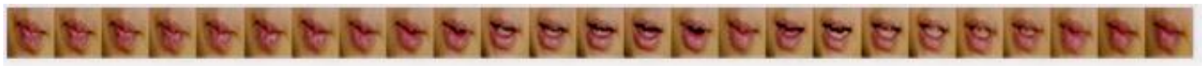


FIGURE 6.3: A sequence of 25 frames of a video of a person saying the phrase "Peace be upon you"

This distinctive transformation, visually captured in Figure 6.4, goes beyond conventional data processing. By concatenating the selected frames into a singular image, our methodology enhances the input structure for proposed models. The resulting 224 x 224 pixel image serves as an optimized representation, enabling our models to effectively exploit the spatial arrangement of mouth frames for enhanced Visual Speech Recognition performance.

FIGURE 6.4: Examples of concatenated sentences.

6.6.2 Data Augmentation

In the realm of deep learning, especially for models like the proposed CNN and ViT, the abundance of high-quality data is paramount to attaining optimal results and preventing

overfitting. Data augmentation, an indispensable facet of the training process, is strategically employed to augment the dataset's richness. Our approach goes beyond conventional augmentation, incorporating nuanced adjustments such as vertical and horizontal shifts and random zoom.

This augmentation strategy serves a dual purpose. Firstly, it introduces variability to the dataset, preventing the model from overfitting to specific patterns and enhancing its ability to generalize across diverse scenarios. Secondly, it promotes robustness, ensuring that the model can effectively handle variations in speech patterns and facial expressions encountered in real-world scenarios.

The combination of concatenation and data augmentation, as meticulously outlined in this section, is a testament to our rigorous approach to dataset optimization. By systematically enhancing the dataset's diversity and richness, we empower our proposed models in Visual Speech Recognition to deliver superior performance across a spectrum of real-world conditions.

6.7 Conclusion

In conclusion, this chapter has provided an exhaustive exploration of the pivotal role played by feature extraction, with a specific focus on Mouth Detection and Tracking, in the domain of Visual Speech Recognition (VSR). We delved into the significance of effective feature extraction as the crucial link between raw visual input and the nuanced information required for accurate model training and subsequent predictions.

The chapter comprehensively covered the concept of Region of Interest (ROI) and its implementation through the robust Dlib library. Additionally, we introduced the YOLO (You Only Look Once) model for real-time mouth detection, elucidating the entire training process, including data preparation, model architecture, hyperparameter tuning, and evaluation metrics. A meticulous comparison between the Dlib and YOLO approaches was presented, aiding researchers in making informed choices based on factors like accuracy, speed, and resource efficiency.

Furthermore, we provided insights into the post-detection data processing steps, encompassing normalization, cleaning, and data augmentation techniques. The pivotal step of concatenating utterances, specifically focusing on cropped mouth regions, was highlighted as a strategy to enhance contextual information for improved model robustness and generalization.

In the context of dataset optimization, the detailed explanation of the concatenation process, transforming temporal data into a spatial format, and the augmentation strategies underscored our commitment to providing high-quality data for superior model performance in Visual Speech Recognition.

This chapter serves as a comprehensive guide for researchers and practitioners in the field, offering practical insights and methodologies for effective feature extraction and data processing in the pursuit of accurate Visual Speech Recognition.

Chapter 7

Proposed Methods

In this chapter, which constitutes the third service within our comprehensive framework, we delve into the details of the different architectures proposed for the task at hand. The diverse set of architectures, including our custom CNN proposal, VGG16, Inceptionv3, and Vision Transformer, are meticulously explained, providing a comprehensive understanding of their design principles and applicability.

7.1 Custom CNN Proposal

Our custom Convolutional Neural Network proposal is meticulously crafted to cater specifically to the demands of our framework, ensuring optimal performance for our task. The architecture is intricately designed to capture spatial hierarchies, expand receptive fields, and effectively extract relevant features crucial for the task at hand. We delve into the details of the network's layers, activation functions, and optimization strategies, providing a comprehensive understanding of the design decisions underpinning its effectiveness. Through transparent insights into its inner workings, stakeholders gain a deeper appreciation of its functionality, facilitating seamless integration and potential enhancements.

7.1.1 Our CNN Model (CustomCNN)

The CNN architecture proposed in this study is specifically engineered for the precise classification of images, leveraging deep learning techniques. Comprised of 16 layers, this architecture incorporates a diverse array of components meticulously designed to optimize its performance, as illustrated in Figure 7.1. This design encompasses a series of convolutional and pooling layers strategically arranged to extract features from input images while reducing dimensionality effectively. Furthermore, the architecture integrates dropout and batch normalization layers to mitigate overfitting, ensuring robustness and generalization capabilities. The final classification process is facilitated by fully connected layers, which consolidate extracted features and provide accurate predictions. This comprehensive design underscores the model's efficacy in handling complex image classification tasks.

The CNN model's fundamental components are its convolutional layers, pivotal for feature extraction. With four convolutional layers, each with specific kernel sizes, feature extraction is targeted effectively. The initial Convolutional Layer includes eight (3x3) kernels, followed by the second layer with 16 (3x3) kernels, the third layer with 32 (4x4) kernels, and the fourth layer with 64 (5x5) kernels. These layers are designed meticulously to extract unique features, allowing the model to learn hierarchical representations efficiently, Table 7.1 summarizes the CustomCNN architecture.

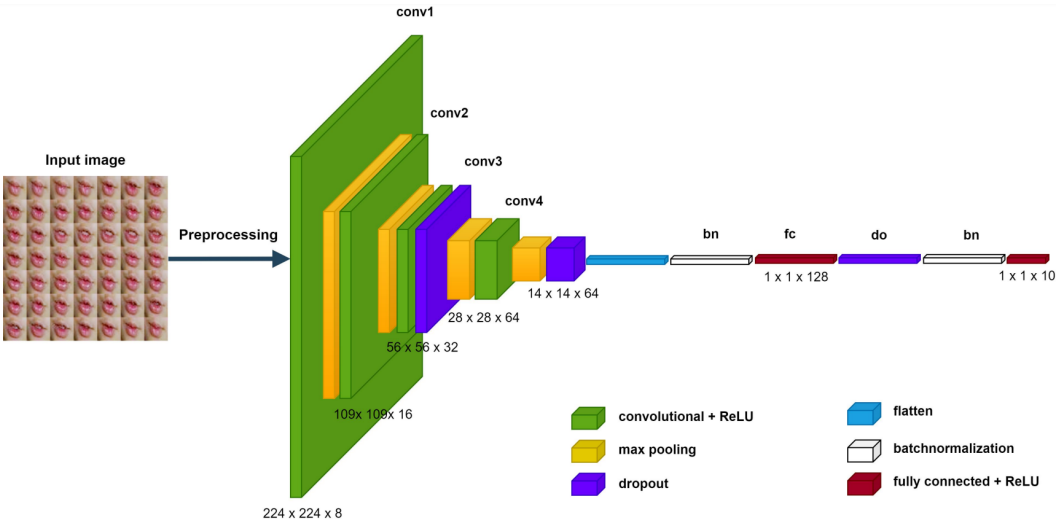


FIGURE 7.1: The architecture of CustomCNN employs a sequence of convolutional and pooling layers to extract features and reduce dimensions. Dropout and batch normalization layers are also incorporated to mitigate overfitting and enhance training efficiency. The final classification is executed by fully connected layers.

In addition, the architecture incorporates six MaxPooling layers with (2x2) kernel sizes to downsample feature maps, preserving crucial information while reducing computational complexity.

To address overfitting, three Dropout layers are integrated into the model, selectively deactivating neurons during training to prevent excessive memorization of the training data.

Moreover, two BatchNormalization layers normalize activations within each batch, facilitating faster training and enhancing generalization performance.

The model's output is produced by dense layers, with the first layer comprising 128 neurons using the ReLU activation function. The second dense layer contains 10 neurons, aligning with the model's classification classes. Finally, the output layer employs the Softmax activation function to normalize the model's output into class probabilities.

In summary, CustomCNN model is proficient in extracting features from images and achieving precise classification. It employs `convolutional layers` to extract features, `MaxPooling layers` to reduce dimensions, `dropout layers` to mitigate overfitting, and `BatchNormalization layers` to normalize activations. Additionally, the architecture integrates dense layers for the final classification, featuring ReLU activation in the initial hidden layer and Softmax activation in the output layer. This model showcases the capability to attain exceptional accuracy in image classification endeavors.

7.2 VGG16

The VGG16 architecture, celebrated for its simplicity and remarkable performance, serves as a cornerstone in our proposed methodologies. In this section, we delve into a comprehensive analysis of the network's structure, which encompasses a series of convolutional and fully

TABLE 7.1: Architecture Details of the CustomCNN Model

Layer Type	Details	Description
Convolutional	1st Layer:	8 filters, size (3,3)
Layers	2nd Layer:	16 filters, size (3,3)
	3rd Layer:	32 filters, size (4,4)
	4th Layer:	64 filters, size (5,5)
MaxPooling	Kernel Size:	(2,2)
Layers	Number of Layers:	6
Dropout	Number of Layers:	3
BatchNormalization	Number of Layers:	2
Dense	1st Layer:	128 neurons, ReLU activation
Layers	2nd Layer:	10 neurons, Softmax activation
	Output Layer:	Softmax activation

connected layers. We meticulously explore the unique attributes that render VGG16 well-suited for the challenging domain of image classification. Additionally, we meticulously outline considerations and strategies for fine-tuning and customizing the pre-trained VGG16 model to seamlessly integrate with our specific application requirements. Through this meticulous examination, we aim to harness the full potential of the VGG16 architecture to achieve optimal performance and efficacy in our image classification tasks.

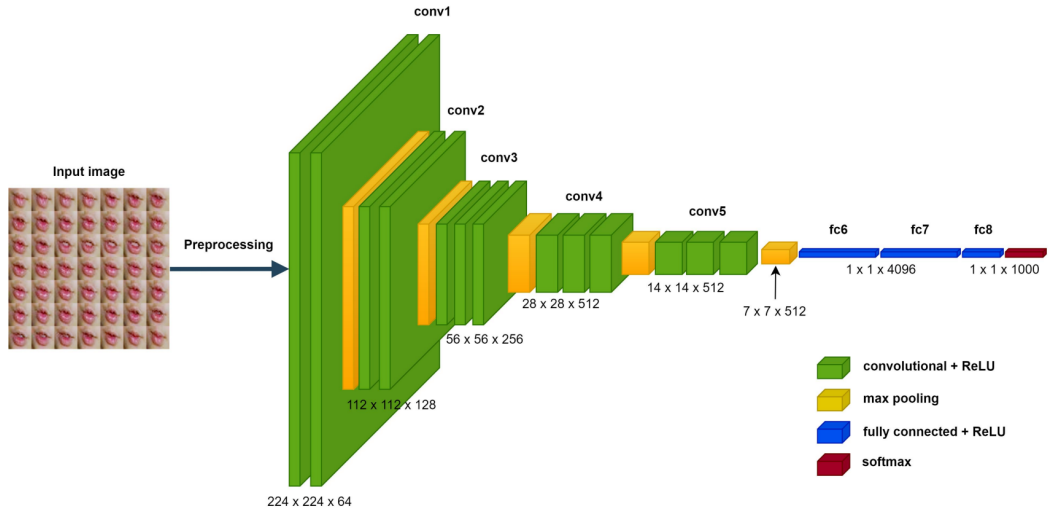


FIGURE 7.2: The VGG16 model we're using for the BBC LRW dataset has a specific structure. It takes in images that are 224 pixels wide, 224 pixels tall, and have 3 color channels. These images go through 13 layers that find important features. We use pre-trained weights from ImageNet to help the model learn faster. After going through all the layers, the final shape of the output is $7 \times 7 \times 512$. Sometimes, we might need to make small adjustments, called fine-tuning, to make sure the model works well for our project's needs.

Architecture VGG16 is characterized by its straightforward architecture, comprising 16 layers:

- The first 13 convolutional layers extract features using 3×3 filters, moving one pixel at a time, and adding padding to keep the output size the same.
- Max-pooling layers shrink the image by half in each direction using 2×2 filters, while still keeping important features.
- Lastly, there are three fully connected layers that help classify the image into different categories.

Deep Learning Strengths VGG16's effectiveness stems from its deep architecture, which enables it to analyze images at multiple levels of complexity. This depth allows the model to discern details, such as textures, shapes, and patterns, at various scales within objects and scenes. By exploring these features across different levels of abstraction, VGG16 can accurately classify images by recognizing both broad categories and fine-grained distinctions. This capability makes it well-suited for tasks requiring nuanced understanding and precise classification, such as image recognition and object detection.

Versatility VGG16's simplicity, widespread availability, and consistently high performance have established it as a preferred option in the research community. Its straightforward design and well-documented performance make it an invaluable benchmark for assessing the efficacy of other deep learning architectures.

Implementation and Accessibility VGG16 offers a notable advantage in its ease of implementation, making it accessible even to researchers with limited resources or expertise. Furthermore, its open-source nature and cost-free availability further contribute to its widespread adoption within the research community.

Fine-Tuning In scenarios where projects have particular demands or limitations, adjusting VGG16's parameters and layers may be essential. This fine-tuning process helps tailor the model's performance to meet the specific needs and circumstances of the project, ensuring optimal results.

Model Structure Figure 7.2 illustrates the architecture of the VGG16 model as applied to the BBC LRW dataset. This depiction offers insight into how the model processes data and the overall structure utilized in this specific context.

- The model analyzes input images sized at 224 x 224 pixels with three color channels, traversing through thirteen convolutional layers for feature extraction.
- Employing transfer learning, the model benefits from pre-trained weights sourced from the ImageNet dataset, aiding in optimizing training efficiency by leveraging existing knowledge.
- Upon completion of processing, the output shape is transformed to a dimension of $7 \times 7 \times 512$, indicating the extracted features' representation.
- Fine-tuning may be necessary to adjust the model's parameters and layers to suit the unique demands and limitations of the project, ensuring optimal performance alignment.

In conclusion, the VGG16 architecture stands out for its simplicity, effectiveness, and versatility, making it a preferred choice for various image classification tasks. Its depth allows it to learn features, enabling recognition across different levels of complexity. Moreover, its widespread implementation and accessibility, along with the potential for fine-tuning, contribute to its appeal among researchers. By providing a robust foundation and flexible structure, VGG16 remains a cornerstone in the field of deep learning, facilitating advancements in image recognition and analysis.

7.3 Inceptionv3

In the pursuit of advancing our visual speech recognition framework, we turn our attention to the InceptionV3 model, renowned for its excellence in image recognition tasks. In this section, we delve into the InceptionV3, exploring its architecture, strengths, and suitability for our framework's objectives. By harnessing the power of InceptionV3, we aim to bolster the accuracy and efficiency of visual speech recognition, ultimately enhancing communication accessibility for individuals with hearing impairments.

7.3.1 InceptionV3 Model Overview

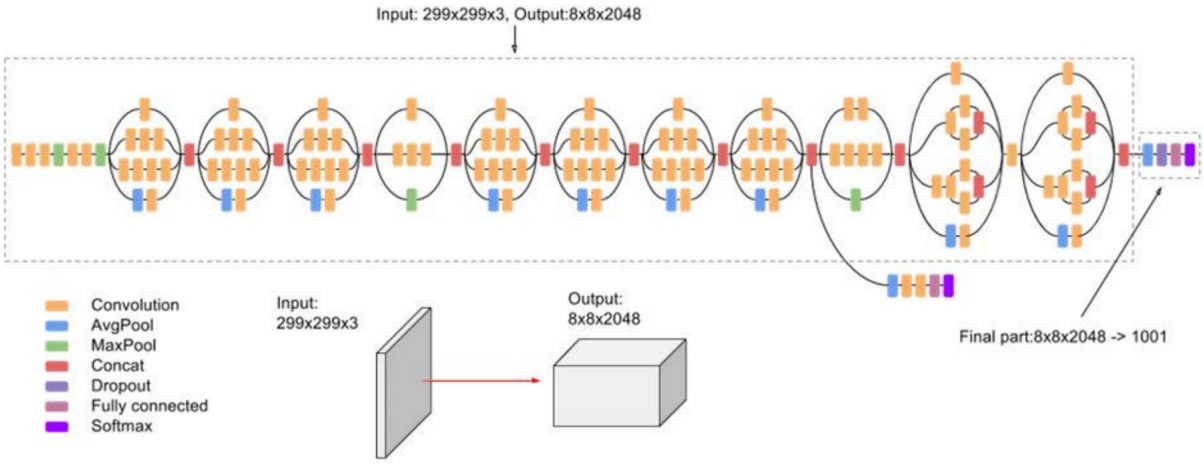


FIGURE 7.3: Diagram outlining the structure of the InceptionV3 model. The model includes convolutional, pooling, and fully connected layers, as well as Inception modules. Arrows illustrate the path of information flow from the input image to the output class probabilities.

Inception-v3 is a deep convolutional neural network (CNN) architecture introduced by Google researchers [100]. It represents a significant advancement over its predecessors, such as AlexNet and VGGNet, by addressing computational efficiency and improving performance in image classification tasks.

- **Inception Module**:
  - The core innovation of Inception-v3 lies in its Inception module, which efficiently captures information at multiple scales within the network.

  - The module achieves this by employing parallel convolutional operations with different kernel sizes (1x1, 3x3, 5x5), pooling operations, and dimensionality reduction techniques, such as 1x1 convolutions, to reduce computational cost.
  - Inception modules are stacked to form the backbone of the architecture, enabling the network to learn hierarchical representations of input images.

- **Factorization Techniques**:
  - Inception-v3 uses factorization techniques to reduce the number of parameters and computational complexity.
  - This includes factorized convolutions, where large convolutional layers are decomposed into smaller convolutions. Auxiliary classifiers are also introduced at intermediate layers to aid in training and regularization.
  - By decomposing large convolutions into smaller ones, the network can capture complex patterns more efficiently while reducing computational overhead.

- **Regularization and Normalization**:
  - Batch normalization is incorporated to accelerate training and improve generalization. Regularization techniques like dropout are used to prevent overfitting and improve model robustness.
  - These techniques help prevent the model from memorizing noise in the training data and improve its ability to generalize to unseen examples.

- **Output Layer**:
  - The architecture concludes with average pooling followed by a softmax layer for classification.
  - Average pooling aggregates spatial information from the final convolutional feature maps and reduces their dimensionality before passing them to the softmax classifier.

- **Performance and Applications**:
  - Inception-v3 has demonstrated state-of-the-art performance on datasets like ImageNet, achieving top-5 error rates below 4%.
  - Its efficient design makes it suitable for deployment on resource-constrained devices and a variety of applications beyond image classification, including object detection, segmentation, and transfer learning tasks.
  - Transfer learning with Inception-v3 involves fine-tuning the pre-trained model on a specific task, leveraging the learned features for improved performance with limited data.

In conclusion, InceptionV3 stands out as a powerful and versatile convolutional neural network, renowned for its computational efficiency and scalability. Its innovative architecture, featuring inception modules, facilitates the extraction of features at multiple scales while minimizing computational complexity. With its remarkable performance in various image-related tasks and real-time applications, InceptionV3 continues to be a popular choice among researchers and practitioners in the field of computer vision.

7.4 Vision Transformer (ViT)

The Vision Transformer, initially conceived for natural language processing, undergoes adaptation for image classification within our framework. We delve into the distinctive methodology of converting images into token sequences, expounding on transformer layers, attention mechanisms, and positional embeddings. This discussion illuminates ViT's ability to grasp spatial relationships within images. Additionally, we meticulously explore strategies for training ViT with image data and fine-tuning its performance to ensure optimal results.

Transformers have become increasingly popular in natural language processing (NLP) [103] [71] thanks to their remarkable performance in tasks like machine translation. This groundbreaking DL architecture, powered by self-attention mechanisms, has emerged as a favored choice over CNNs and LSTM models, prompting a growing focus from researchers in this field.

Furthermore, the Transformer architecture has found application in computer vision tasks with the introduction of the ViT proposed in the seminal paper "An Image is Worth 16x16 Words" [24]. ViT leverages the self-attention mechanism inherent in Transformers to tackle image classification. It achieves this by initially segmenting images into sequences of flattened patches, which are then fed into a standard Transformer encoder along with positional embeddings. This process enables ViT to effectively extract pertinent features from the image. Notably, ViT has showcased outstanding performance across various computer vision benchmarks, encompassing tasks such as image classification, object detection, and semantic image segmentation.

The methodology proposed by Dosovitskiy et al. [24] presents a detailed approach for training Vision Transformer (ViT). Initially, the image is segmented into patches, which are then flattened to generate lower-dimensional linear embeddings. Next, positional embeddings are incorporated into these flattened patches. Subsequently, a standard Transformer encoder processes the resulting sequence. The model undergoes pre-training using image labels in a fully supervised manner on a large dataset. Finally, the pre-trained image classification model is fine-tuned on the downstream dataset. These systematic procedures have enabled ViT to achieve exceptional performance across various computer vision tasks, establishing it as a competitive technique (see Figure 7.4) [117].

7.4.0.1 Preprocessing stage:

During this stage, the inputs undergo preparation for the encoder, which requires a vector sequence as input. Given that our data consists of 2D images, these images are divided into non-overlapping patches to form a sequence akin to words. These patches are then transformed into one-dimensional vectors to create an embedding sequence, with positional encoding ensuring spatial information retention. Additionally, a learnable classification

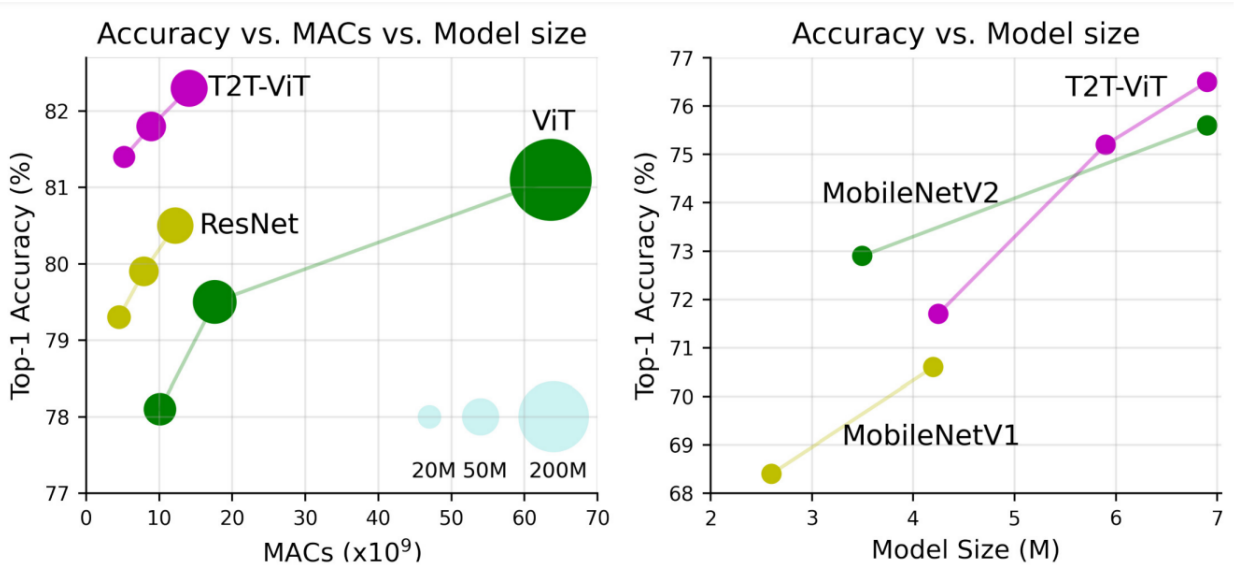


FIGURE 7.4: A comparative analysis was carried out between T2T-ViT, ViT, ResNets, and MobileNets, all trained anew on the ImageNet dataset. On the left, the performance curve illustrates the relationship between MACs and top-1 accuracy, while on the right, the performance curve depicts the association between model size and top-1 accuracy.

token (referred to as the *cls* token) is appended at the beginning of the embedding sequence, serving as a reference point for classification tasks later on.

In our approach, images are reshaped into $\mathsf{I} \in \mathbb{N}^{\mathsf{whc}}$, where w and h represent the width and height of the images respectively, and c denotes the number of channels. For instance, the shape is *(w, h, c) = (224, 224, 3)*. Each image I is then divided into patches $\mathsf{I_p} \in \mathbb{N}^{\mathsf{np^2c}}$, where p signifies the patch size and n denotes the number of patches, calculated as $\mathsf{n} = \frac{\mathsf{wh}}{\mathsf{p^2}}$. Consequently, each image yields $\frac{224*224}{32^2} = 49$ patches, representing various mouth regions extracted from video frames of a sentence. These patches are flattened, and a *cls* token $\mathsf{T_{cls}}$ is added at the sequence's onset. Additionally, position embeddings $\mathsf{E_{pos}}$ are incorporated to preserve the patch order, resulting in the sequence of embedding vectors $\mathsf{z_0}$ of dimension d (refer to Eq 7.1), which serves as input for the subsequent encoder stage.

$$\begin{aligned} \mathsf{z_0} = [T_{cls} + I_p^i E + E_{pos}], \qquad & i = 1, \ldots, n, \\ & E \in \mathcal{N}^{3p^2 d}, \\ & E_{pos} \in \mathcal{N}^{(n+1)d} \end{aligned} \tag{7.1}$$

7.4.0.2 Encoder stage

In the encoder stage, the aggregation of information from the entire input sequence occurs. This process involves stacking a specific number L of identical encoders, each containing four key components: a normalization layer (NL), Multi-head self-attention (MHSA), a feed-forward layer (MLP), and a residual connection shortcut. The outputs of each encoder serve as the inputs to the subsequent encoder, facilitating the flow of information through the model.

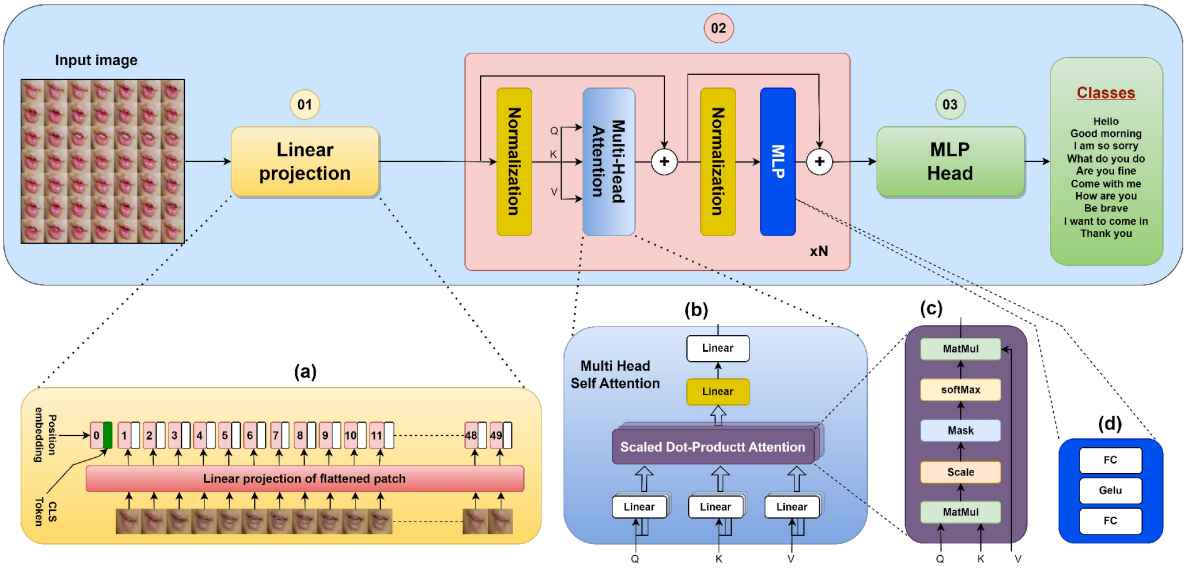


FIGURE 7.5: The Vision Transformer architecture is structured into three primary parts. The initial phase, termed the embedding part (1), integrates a linear embedding component (a). Following this, the encoding part (2) employs the Transformer encoder module, which comprises two pivotal components: Multi-head self-attention (MHSA) (b) and the self-attention block (SAB) head (c). Lastly, the classification part (3), represented by (d), involves a fully connected layer.

The attention mechanism facilitates the extraction and aggregation of information from all inputs, enabling the transformer to achieve rapid parallel processing. Specifically, upon arrival of z_0 in the encoder, it undergoes normalization by LN before being processed by MHSA. The resulting output $\mathsf{Z}_\mathsf{l}^{'}$ is calculated according to Eq. 7.2.

$$\mathsf{Z}_\mathsf{l}^{'} = \mathsf{MHSA}(\mathsf{LN}(\mathsf{Z}_{\mathsf{l}-1})) + \mathsf{Z}_{\mathsf{l}-1}, \mathsf{l} = 1, ..., \mathsf{L} \tag{7.2}$$

Multi-headed self-attention comprises a fixed number of self-attention blocks n, with each self-attention block (SAB) dedicated to discerning dependencies between each patch and others in the sequence. MHSA ensures multiple perspectives by employing multiple SABs, enabling the model to capture diverse viewpoints across the image. As depicted in parts (b) and (c) of the global ViT architecture in Figure 7.5, the self-attention structure facilitates understanding the contextual relationships between patches. This process allows ViT to analyze potential contextual associations for each element within the input sequence processed by the SAB.

- To create the query (Q), key (K), and value (V) vectors, we multiply the embedding vector (E) with three learnable matrices ($\mathsf{W}^\mathsf{Q}, \mathsf{W}^\mathsf{K}, \mathsf{W}^\mathsf{V}$) whose weights are adjusted during the training process. Mathematically, this can be expressed as:

$$\mathsf{Q} = \mathsf{E} \cdot \mathsf{W}^\mathsf{Q}$$

$$\mathsf{K} = \mathsf{E} \cdot \mathsf{W}^\mathsf{K}$$

$$\mathsf{V} = \mathsf{E} \cdot \mathsf{W}^\mathsf{V}$$

where Q, K, and V represent the query, key, and value vectors respectively, and $\mathsf{W^Q}, \mathsf{W^K}, \mathsf{W^V}$ denote the learnable matrices.

- To calculate the dot product between the Q and K vectors ($\mathsf{QK^T}$), we simply multiply the Q vector by the transpose of the K vector. Mathematically, this can be expressed as:

$$\mathsf{QK^T} = \mathsf{Q} \cdot \mathsf{K}^T$$

  This operation computes the pairwise dot product between the query and key vectors, resulting in a matrix where each element represents the relevance or similarity score between the corresponding elements in the sequences, more precisely in our case, determining the relevance between the treated mouth frame and the other frames.

- To normalize the obtained score, we scale it by dividing by the square root of the dimensions of the key vector ($\sqrt{\mathsf{d^K}}$). This scaling ensures that the gradients remain stable during training. Mathematically, the normalized score $\mathsf{S}^{'}$ is calculated as:

$$\mathsf{S}^{'} = \frac{\mathsf{QK^T}}{\sqrt{\mathsf{d^K}}}$$

  Where $\mathsf{d^K}$ represents the dimensions of the key vector.

- To translate the scores into weights, we apply the Softmax activation function to the scaled scores obtained from the previous step. Mathematically, this is expressed as:

$$\mathsf{S} = \mathsf{Softmax}(\mathsf{S}^{'})$$

  The Softmax function normalizes the scores, ensuring they are all positive and sum up to 1, effectively creating a probability distribution over the elements in the sequence.

- The output of the self-attention block (SAB) is produced by multiplying the value vector of the current element by the output of the Softmax function. This operation ensures that irrelevant elements are discarded, and only the relevant ones are retained.

The equation below Eq7.3 summarizes what happens in the i-th self-attention block:

$$\mathsf{head_i} = \mathsf{Attention}(\mathsf{Q_i}, \mathsf{K_i}, \mathsf{V_i}) = \mathsf{Softmax}(\frac{\mathsf{Q_i K_i^T}}{\sqrt{\mathsf{d^K}}})\mathsf{V_i}, \mathsf{i} = 1, ..., \mathsf{h} \tag{7.3}$$

Each SAB generates h new embeddings, with each vector subsequently concatenated by the MHSA mechanism, ensuring the preservation of the desired shape through learnable projection weights W, as described in Eq. 7.4, where h represents the number of self-attention heads.

$$\mathsf{MHSA}(\mathsf{Q}, \mathsf{K}, \mathsf{V}) = \mathsf{Concat}(\mathsf{head_1}, ..., \mathsf{head_h})\mathsf{W} \tag{7.4}$$

The resulting embeddings of the sequence $\mathsf{Z_l^{'}}$ are then passed through a feed-forward network, as indicated in the subsequent equation (Eq. 7.5).

$$\mathsf{Z_l} = \mathsf{MLP}(\mathsf{LN}(\mathsf{Z_l^{'}})) + \mathsf{Z_l^{'}}, \mathsf{l} = 1, ..., \mathsf{L} \tag{7.5}$$

7.4.0.3 Classification stage

Upon completing the encoding of the input image through L encoders, we arrive at the final stage. Here, the process involves extracting the first element, the [cls] token, and retaining the last representation while discarding other elements in the sequence. This chosen representation is then forwarded to an external classifier, typically realized as a Multi-Layer Perceptron (MLP). The network's ultimate output is a vector of shape (1, `numClass`), containing the probabilities associated with each class within the `numClass` classes. This procedure can be mathematically expressed as:

$$\mathsf{Y} = \mathsf{LN}(\mathsf{Z}_{\mathsf{l}}^{0})$$

Here, Y represents the final output of the classification stage, and $\mathsf{Z}_{\mathsf{l}}^{0}$ is the output of the last encoder layer.

7.5 Experimental Results and Discussion

Within the Experimental Results and Discussion section, our primary aim was to set a baseline performance for the BlidAVS10 dataset. Additionally, we introduced deep learning models that utilize both Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) technology. These models were evaluated through comparisons involving pre-training and training from scratch methodologies.

7.5.1 Experiments Objective

The experiments aim to (i) assess the performance of deep learning models, encompassing both CNNs and ViT models, on the newly introduced dataset. (ii) Examine the influence of pre-trained and scratch-trained variations. (iii) Explore the potential benefits and constraints of employing color and grayscale representations with CustomCNN for this particular dataset. Lastly, (iv) compare the efficacy of CNN and ViT models to ascertain the optimal approach for the dataset at hand.

7.5.2 Experimental Setup

All experiments were conducted on our dataset comprising 1383 videos. We utilized CNNs and ViT models, with hyperparameters configured as outlined in Table 7.2.

The experiments were performed on a desktop system equipped with high-performance hardware components, including an NVIDIA GeForce RTX 3060 Ti graphics processing unit (GPU) with 8 gigabytes of memory, an AMD Ryzen 7 5800X 8-Core central processing unit (CPU), and 16 gigabytes of random access memory (RAM). These hardware selections were made for their ability to efficiently handle deep learning tasks, ensuring rapid and effective processing. The robust performance of these components facilitated the execution of computationally intensive experiments involving both CNNs and ViT technology. Leveraging such advanced hardware resources enabled us to conduct experiments with optimal efficiency, ensuring the generation of accurate and dependable results.

7.5.3 Results

In this section, we present the outcomes of our experiments conducted on our dataset. Table 7.3 provides a summary of the performance metrics for the various models utilized,

Table 7.2: Main hyperparameters for each experiment; SCCE= sparse Categorical Cross entropy, CCE= categorical cross entropy.

Model	Optimizer	Loss	Learning Rate	Pretrained	Epochs	Early Stopping
Vit	Custom	SCCE	Custom	from scratch / vitbase-patch 32-224-in21k	30	7
CustomCNN	adam	SCCE	lr=0.001 momentum=0.99	From Scratch	70	7
InceptionV3	SGD	CCE	lr=0.0001 momentum=0.90	Imagenet	70	7
VGG16	SGD	CCE	lr=0.0001 momentum=0.90	Imagenet	200	7

showcasing their training and validation accuracy percentages. These findings offer valuable insights into the efficacy of each model within the scope of our investigation.

Table 7.3: The results of different models we employed

Methods	Train %	Validation %
CustomCNN grayscale	96	90
CustomCNN	90	85
InceptionV3	83	79
VGG16	80	80
ViTransformer (transfer learning)	99	98
ViTransformer (from scratch)	-	-

7.5.3.1 Evaluation of Model Performance Experiment

In our preliminary experiment, we developed deep learning models tailored to the `BlidAVS10` dataset to generate captions corresponding to the videos. Initially, we utilized a training subset from our dataset to derive lip movement models. Following this, we evaluated the performance of each deep learning model by assessing its accuracy.

- **Evaluation of CustomCNN Performance on Color Images** For the assessment involving CustomCNN on color images, our objective was to analyze the model's efficacy in accurately capturing lip movements while considering the presence of color information.
  - *Objective:* The aim was to evaluate CustomCNN's performance when processing colored inputs and its capability to extract relevant features from the varied visual information present in lip movements.
  - *Analysis:* The results depicted in Figure 7.8 illustrate the accuracy (see Figure 7.6) and loss (refer to Figure 7.7). Notably, there is a significant improvement in validation accuracy, reaching 85% on the validation set. This indicates that CustomCNN adeptly utilizes color information, thereby enhancing its ability to comprehend lip movements.

- **Performance of CustomCNN on Grayscale Images** The objective was to evaluate how CustomCNN performs when processing grayscale inputs and its capability to extract relevant features from grayscale lip movement data.

  - *Objective:* The aim was to gauge the model's performance in the absence of color information, accentuating its capacity to depend solely on intensity fluctuations for accurate lip movement recognition.

  - *Analysis:* The outcomes depicted in Figures 7.11 reveal the accuracy (Figure 7.9) and loss (Figure 7.10) metrics for CustomCNN when applied to grayscale images. Interestingly, the accuracy demonstrates a further enhancement, reaching 90%. This underscores the model's resilience and indicates its adeptness in extracting crucial features even in the absence of color data.

- **Performance of InceptionV3:** In this experiment, we scrutinized the efficacy of the InceptionV3 model relative to other CNN-based approaches.

  - *Objective:* The primary aim was to evaluate the ability of the InceptionV3 architecture to capture spatial features in lip movements.

  - *Analysis:* Figures 7.20 and Table 7.3 illustrate the accuracy (see Figure 7.18) and loss (see Figure 7.19) results. While InceptionV3 demonstrates reasonable performance, the findings suggest that CustomCNN models, particularly in grayscale, outperform InceptionV3, achieving a validation accuracy of 79%. This underscores the significance of our proposed architecture.

- **Performance of VGG16** In this experiment, we assessed the efficacy of the VGG16 model, aiming to compare its performance with other CNN-based algorithms in the context of lip movement recognition.

  - *Objective:* The objective was to evaluate the capability of VGG16 in capturing features within lip movements and to assess its suitability for our specific dataset.

  - *Analysis:* Figures 7.17 and Table 7.3 display the accuracy (Figure 7.15) and loss (Figure 7.16) results. While VGG16 demonstrates reasonable performance, particularly in terms of training accuracy, the model's validation accuracy slightly lags behind CustomCNN models and ViT. This suggests that, within the scope of our study, CustomCNN and ViT models may be preferable due to their superior generalization ability.

- **Performance of ViT** The experiment on the ViT focused on evaluating the model's performance and the impact of transfer learning.

  - *Objective:* The objective was to explore the ViT model's capacity in capturing long-range dependencies within lip movements and to evaluate the efficacy of transfer learning.

– *Analysis:* Figures 7.14 and Table 7.3 present the accuracy (Figure 7.12) and loss (Figure 7.13) results. Remarkably, the ViT model with transfer learning achieves an outstanding 98% validation accuracy, indicating its superiority. These findings suggest that transfer learning significantly enhances ViT's performance, rendering it a robust choice for lip movement recognition tasks.

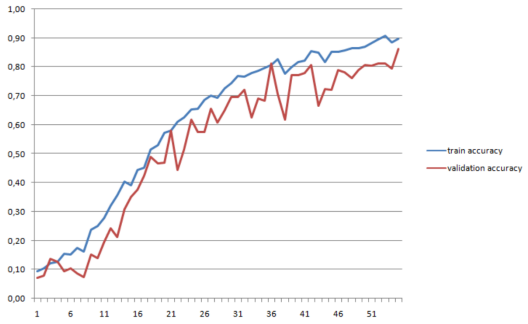


FIGURE 7.6: Accuracy of the CNN proposed results on color images.

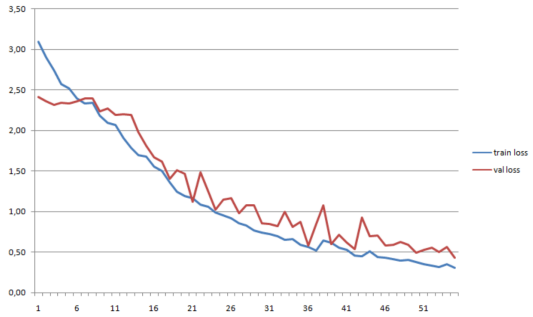


FIGURE 7.7: Loss of the CNN proposed results on color images.

FIGURE 7.8: Accuracy and loss results obtained for CustomCNN on colored images.

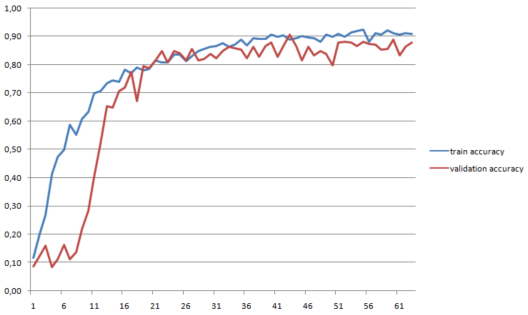


FIGURE 7.9: Accuracy of the CNN proposed results on grayscale images.

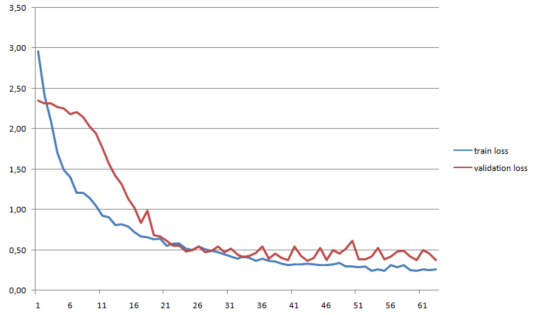


FIGURE 7.10: Loss of the CNN proposed results on grayscale images.

FIGURE 7.11: Accuracy and loss results obtained for CustomCNN on grayscale images.

With validation accuracy of 90% and 98%, respectively, it is clear that CustomCNN models and the ViT model with transfer learning perform better than other models. This exemplifies how the framework presents and enhances the ability to extract extremely subtle lip movement expressions. The user has an effective tool to generate a voice or caption related to the required video for numerous jobs by utilizing the `BlidAVS10` framework.

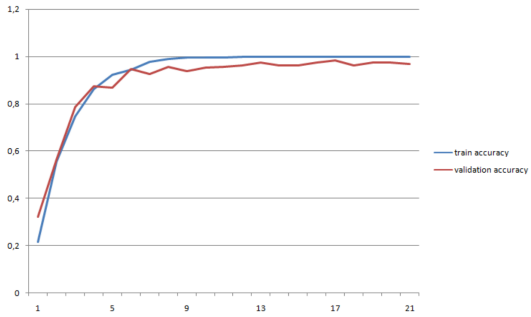


FIGURE 7.12: Vit algorithm accuracy results

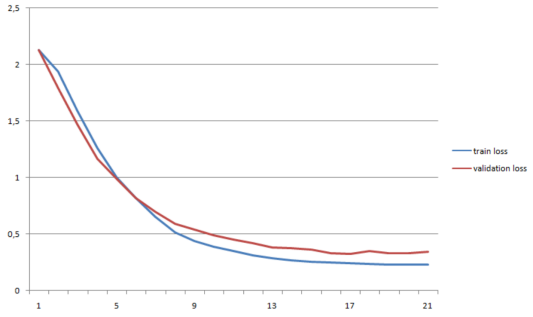


FIGURE 7.13: Vit algorithm loss results

FIGURE 7.14: Accuracy and loss results obtained for the Vit algorithm using CNN-based methods

7.5.3.2 Effects of the Experiment with Scratch-Trained Variants and Pre-trained

The objective of our study was to assess the effectiveness of our models in the proposed Arabic Visual Speech Recognition (VSR) framework, BlidAVS10. We aimed to investigate the impact of two different training methodologies: training from scratch and utilizing transfer learning.

CustomCNN and ViT models were trained entirely from scratch, while models like VGG16, InceptionV3, and another ViT model underwent training through transfer learning. The results, as presented in Table 7.3, demonstrate that the ViTransformer model with transfer learning consistently outperformed other variants, achieving an impressive validation accuracy of 98%. This emphasizes the robustness and effectiveness of transfer learning, particularly in the context of the ViT architecture.

Additionally, it is worth noting that CustomCNN models performed significantly better than VGG16 and InceptionV3 in terms of validation accuracy when trained from scratch. This highlights the advantage of our approach, especially when initializing models without pre-trained weights.

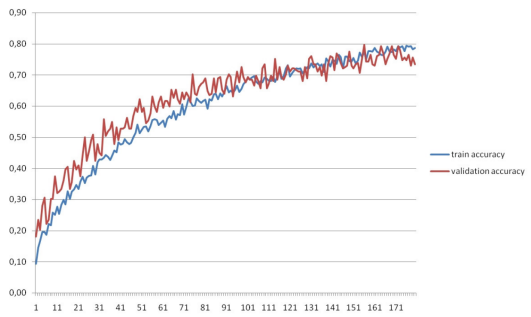


FIGURE 7.15: VGG16 accuracy results

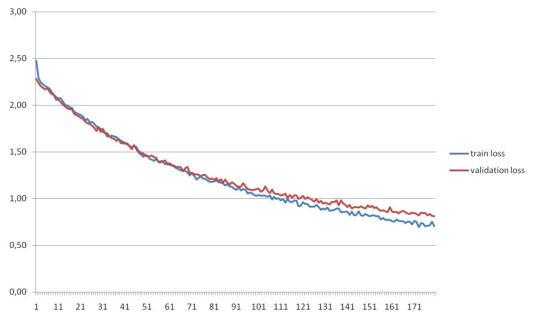


FIGURE 7.16: VGG16 loss results

FIGURE 7.17: Accuracy and loss results obtained for VGG16 using CNN-based algorithms

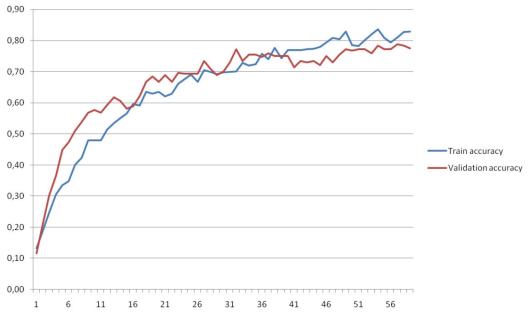


FIGURE 7.18: InceptionV3 accuracy results

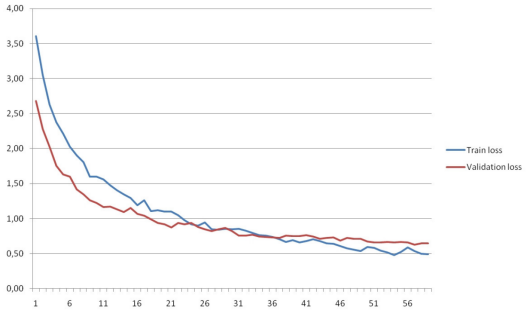


FIGURE 7.19: InceptionV3 loss results

FIGURE 7.20: Accuracy and loss results obtained for InceptionV3 using CNN-based algorithms

7.5.3.3 Experiment with Color and Grayscale Representations

The experiment conducted to explore the benefits and limitations of using color and grayscale representations in CustomCNN for the dataset yielded interesting results. By training CustomCNN models with both color and grayscale inputs, we were able to compare their performance. The summarized outcomes, as shown in Figure 7.21, clearly indicate that the grayscale variant of CustomCNN achieved the highest validation accuracy, an impressive 90%. This surpassed the performance of the color variant and other models in our evaluation. These findings suggest that grayscale representations offer distinct advantages for the specific task within the dataset, potentially due to the reduced complexity and improved feature extraction capabilities in the absence of color information.

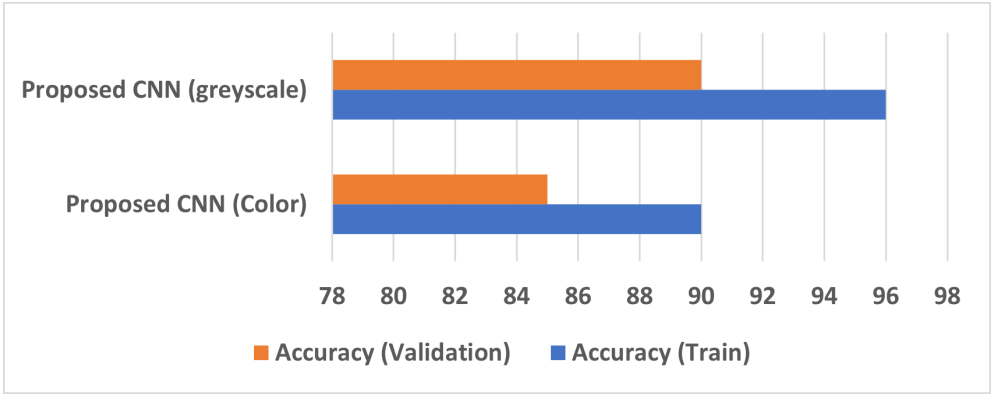


FIGURE 7.21: Comparison of Accuracy Between CustomCNN Models Using Color and Grayscale Representations

The findings from this experiment provide valuable insights into optimizing model performance, particularly emphasizing the suitability of grayscale representations for this specific dataset. This knowledge can guide researchers in future endeavors related to BlidAVS10 research, helping them adopt more effective approaches for visual speech recognition tasks. The inherent simplicity of grayscale images, with only one channel compared to the three channels (red, green, and blue) in color images, reduces the overall complexity of the input data. This reduction in complexity may have allowed the CNN to

focus on crucial features and patterns in lip movements without being influenced by color variations. The absence of color-related noise could have contributed to the improved performance, highlighting the advantages of utilizing grayscale representations for tasks that require precise feature extraction and recognition.

Additionally, the success of the grayscale variant underscores the adaptability of the BlidAVS10 dataset. It demonstrates the effectiveness of our approach and reinforces the potential for further advancements in visual speech recognition. By leveraging grayscale representations, researchers can harness the power of simplified input data to enhance the performance and accuracy of their models. This experiment serves as a stepping stone towards developing more robust and efficient systems for visual speech recognition, ultimately benefiting various applications in the field.

7.5.3.4 Experiment on Comparing the Performance of CNN and ViT Models

The performance of the CNN models and the ViT model was thoroughly compared in our final experiment to determine the most appropriate approach for our dataset. The results of this comparative analysis are visually presented in Figure 7.22. The findings of the experiment highlight the remarkable performance of the ViT model with transfer learning, surpassing CustomCNN models in terms of validation accuracy with an impressive rate of 98%. Additionally, the evaluation of loss functions further emphasizes the effectiveness of the ViT model in minimizing loss.

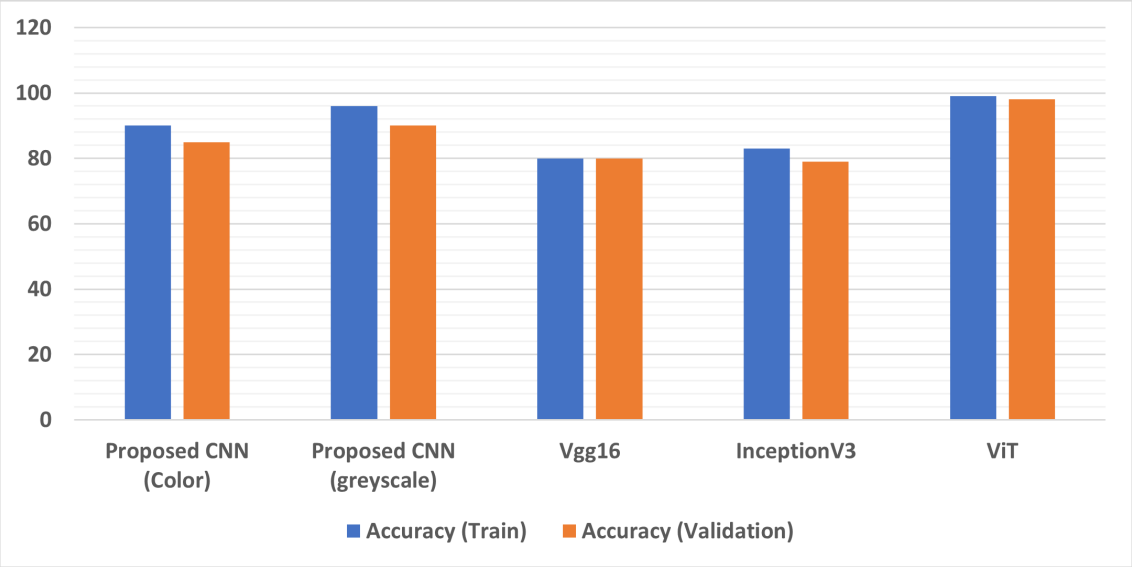


FIGURE 7.22: Comparison of Accuracy Across Deep Learning Models on Training and Validation Datasets

To summarize, the results indicate that both the ViT model with transfer learning and the grayscale variant of CustomCNN models have demonstrated exceptional performance, showcasing their suitability for the dataset. In comparison to other models utilized in this study, the ViT model exhibited exceptional accuracy on both the training and validation sets, converging rapidly. Furthermore, the validation accuracy closely mirrored the training accuracy, indicating that the model successfully acquired informative data representations and could generalize effectively to new data.

Therefore, based on these findings, it can be concluded that CustomCNN and ViT models may be preferred over the VGG16 and InceptionV3 models due to their superior generalization ability. The ViT model, in particular, showcased outstanding performance and proved to be highly effective in handling our dataset. These results provide valuable insights for future research and applications in the field of computer vision.

7.5.4 Discussion

The results of the experiment indicate that by utilizing the ViT approach through transfer learning and pre-trained models, a significant improvement in accuracy can be achieved. During the training phase, the accuracy reached an impressive 99%, while during validation, it achieved 98%. However, when attempting to train the model from scratch, this approach did not yield as favorable outcomes. On the other hand, the CNN approach demonstrated remarkable performance, surpassing a 90% accuracy rate during validation and 96% during training, with a loss function value of 0.3. Interestingly, this was accomplished by training the model from scratch using a grayscale image database, which proved to be more effective than employing color images with three channels. Moreover, the proposed model outperformed previously trained models such as InceptionV3 and VGG16 in terms of both accuracy and loss function value.

The results of this study indicate that the ViT approach is better suited for tasks that have pre-trained models readily available. On the other hand, the CNN approach may be more advantageous for tasks that involve smaller datasets or datasets that are specific to a particular domain. Moreover, training models using grayscale images instead of color images with three channels could potentially offer benefits, as grayscale images are easier to train and may produce more precise outcomes. In conclusion, these findings offer valuable insights for professionals and researchers working in the areas of computer vision and deep learning.

The anticipated outcome of the ViT pre-trained method surpassing the scratch-trained alternative was not surprising, considering the Transformer's dependence on abundant data for successful training. Hence, the results obtained in this aspect are in line with what was expected. On the other hand, the experiment emphasized that customized models built on CNNs exhibited superior performance when trained from scratch, outperforming those utilizing standard pre-trained models. This emphasizes the superiority of tailored models over generic ones.

Besides the factors previously addressed, the ViT approach's superiority can be attributed to two more factors. Firstly, it utilizes a visual speech representation technique that involves concatenating frames of the region of interest into a single image. This allows the ViT approach to capture temporal information more effectively by preserving the sequential order of the frames.

Additionally, the underlying philosophy of each method differs. The CNN methodology adopts a local perspective, emphasizing the identification and preservation of significant and representative features within the image while disregarding irrelevant details. This is achieved through the utilization of filters that discern the interdependencies among neighboring pixels, with diverse filter sizes ensuring the recognition of various dependencies within the pixel patch. While this strategy enables CNNs to concentrate solely on crucial features, it may impede the detection of temporal information inherent in the sequential arrangement of frames within the region of interest. In contrast, the ViT methodology adopts a holistic perspective of the entire image, partitioning it into patches and establishing

connections among all patches. This empowers the ViT model to more efficiently comprehend the temporal information embedded within the input image, presenting an advantage over the CNN model in this specific context. By considering these aspects, practitioners can make well-informed decisions regarding the most suitable approach for their specific task.

The results of the experiment confirm the effectiveness and feasibility of the recently created dataset in building automated VSR systems. Additionally, the suggested methods demonstrate impressive performance in lip-reading, with the ViT approach being particularly outstanding. These discoveries validate the efficiency of both the new dataset and the proposed methods, indicating their dependability and appropriateness for utilization in the advancement of automated VSR systems. The outcomes of the study emphasize the potential of the ViT approach, positioning it as a promising choice for practitioners seeking to improve their lip-reading abilities.

This paper introduces a novel approach aimed at enhancing Visual Speech Recognition (VSR) systems within Arabic audiovisual applications. Our framework, `BlidaAVS10`, is meticulously crafted to elevate the efficiency of VSR systems, distinguished by its originality. To our knowledge, no dataset as comprehensive and specialized as `BlidaAVS10` exists to support automatic Arabic VSR systems, thus bridging a crucial gap in the field and providing a valuable resource for scientists and researchers. Our objective is to inspire the computer vision community to address the challenges associated with conceptualizing and articulating student motivation states, presenting them in an abstract yet machine-readable format.

7.6 Conclusion

This chapter has meticulously outlined the proposed methods for Visual Speech Recognition (VSR), focusing on the integration of custom Convolutional Neural Network, VGG16, Inceptionv3, and ViT architectures within our comprehensive framework.

The custom CNN model has been tailored to address the unique requirements of VSR, emphasizing feature extraction and accurate classification. The incorporation of convolutional layers, MaxPooling layers, dropout layers, and batch normalization layers showcases a holistic approach to visual speech analysis.

VGG16, known for its simplicity and effectiveness, seamlessly fits into our VSR methods. The depth of its architecture enables the learning of complex visual speech features, making it a reliable choice for VSR tasks. Fine-tuning considerations and the utilization of pre-trained weights from ImageNet highlight its adaptability.

Inceptionv3, with its inception modules and focus on feature diversity, contributes a powerful option to our VSR framework. The architecture's capability to handle varying scales of visual speech features and its efficiency in rapid data processing underscore its suitability for VSR tasks.

The Vision Transformer, originally designed for natural language processing, is adeptly adapted for visual speech classification in our framework. Its unique approach of transforming visual speech frames into sequences of tokens, coupled with the self-attention mechanism, demonstrates exceptional performance in various VSR tasks.

In conclusion, the methodologies outlined in this chapter provide a robust foundation for Visual Speech Recognition within our framework. The diverse set of proposed architectures offers flexibility and choice, empowering users to select the approach that best aligns with their specific visual speech recognition needs and objectives.

Part IV

Conclusion and Perspectives

Chapter 8

General Conclusion

8.1 Conclusion

In conclusion, this research endeavors to advance the field of Visual Speech Recognition (VSR) by addressing key challenges and exploring innovative approaches. Through a comprehensive literature review, it is evident that the accurate interpretation of visual cues, particularly lip movements, is crucial for enhancing the performance of VSR models.

8.1.1 Summary of Achievements

This research has achieved significant milestones in the field of Visual Speech Recognition (VSR), encompassing various services within the framework. The exploration commenced with the creation of a comprehensive Arabic dataset, laying the foundation for subsequent services. The dataset creation process involved meticulous planning, equipment selection, and addressing challenges such as participant consent.

Moving forward, the study integrated traditional methods, utilizing the Dlib library for reliable mouth detection within the VSR system. Additionally, the research seamlessly incorporated modern real-time techniques by adopting the YOLO (You Only Look Once) model, focusing on its efficiency in enhancing mouth detection capabilities. A rigorous comparative analysis was conducted between the reliability of traditional methods like Dlib and the real-time efficiency of YOLO.

Transitioning to the recognition phase, the study leveraged Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) approaches. This strategic shift towards expressive deep learning architectures showcased the adaptability of the VSR system, contributing to its overall capabilities.

The development of a systematic and integrated framework for Arabic VSR, covering data collection, processing, feature extraction, and Arabic dataset creation, stands out as a noteworthy achievement. This framework accommodates diverse methodologies and encapsulates the integration of traditional and modern approaches seamlessly.

8.1.2 Contributions of the Thesis

This thesis makes substantial contributions across multiple dimensions. The creation of a comprehensive Arabic dataset addresses a critical gap in Arabic VSR research, facilitating studies in Arabic-speaking contexts. The adoption of Dlib and YOLO for mouth detection showcases a pragmatic approach, where each method contributes distinctively—Dlib for its reliability in traditional methods and YOLO for its efficiency in real-time techniques. This

dual-method strategy ensures a comprehensive and versatile approach to mouth detection within the VSR system.

Moving to the recognition phase, the incorporation of CNNs and ViT represents a pivotal contribution. This strategic shift towards expressive deep learning architectures highlights the adaptability of the VSR system. The thesis emphasizes the significance of embracing diverse approaches, from traditional methods to modern deep learning techniques, within a unified framework.

The developed framework serves as a comprehensive guide for both novice researchers and seasoned practitioners venturing into the field of Visual Speech Recognition (VSR). Its unique strength lies in the seamless integration of both traditional and modern approaches, fostering a balanced and standardized methodology within the VSR domain. The meticulous examination of research questions and the conducted comparative analysis, specifically between Dlib and YOLO for mouth detection and CNNs and ViT for the recognition part, imparts nuanced insights into the multifaceted landscape of visual speech recognition. This research framework not only addresses critical questions but also contributes to advancing the standard practices and methodologies in the dynamic field of VSR.

In summary, this research advances VSR by creating a comprehensive Arabic dataset, adopting a pragmatic approach for mouth detection, and leveraging expressive deep learning architectures for recognition. The developed framework and insights contribute to the ongoing evolution of VSR methodologies.

8.2 Perspectives

8.2.1 Implications of the Study

The outcomes of this research carry substantial implications, resonating across both academic and industrial landscapes. The creation of a comprehensive Arabic dataset not only addresses a significant void in Arabic Visual Speech Recognition (VSR) research but also unlocks potential avenues for cross-cultural studies and diverse applications in various domains. The strategic adoption of a dual-method approach to mouth detection, specifically incorporating the state-of-the-art YOLO (You Only Look Once) model, contributes to a versatile and robust foundation for real-world VSR systems. This approach strikes a crucial balance between the reliability inherent in traditional methods and the efficiency demanded by contemporary, real-time techniques.

8.2.2 Expanding the Dataset

One of the pivotal steps in enhancing the effectiveness and generalizability of VSR models involves expanding the dataset. By incorporating more diverse speakers and a wider array of phrases, the model's ability to generalize across different accents, dialects, and speech patterns is significantly improved. This diversity ensures that the model can effectively handle variations in speaker identity, age, gender, and speaking styles, thereby increasing its robustness and applicability in real-world scenarios. Furthermore, expanding the dataset to include different environmental conditions, such as varied lighting and background settings, can enhance the model's adaptability and reliability under diverse operational contexts.

8.2.3 Multimodal Data Integration

Integrating multimodal data, combining both visual and audio inputs, represents a significant advancement in VSR technology. The fusion of these modalities allows for more comprehensive feature extraction and provides a richer context for speech recognition. This multimodal approach leverages the strengths of each data type, where visual data can aid in deciphering speech in noisy environments, and audio data can provide clarity when visual cues are ambiguous. By synergizing these data sources, the model's recognition performance is markedly enhanced, leading to more accurate and reliable outcomes.

8.2.4 Model Optimization

Continuous model optimization is crucial for achieving superior real-time performance and accuracy in VSR systems. This involves refining deep learning models through techniques such as hyperparameter tuning, architecture redesign, and the implementation of advanced optimization algorithms. Focusing on reducing computational overhead without compromising accuracy is essential for deploying VSR systems in real-time applications. Moreover, optimizing models to ensure they perform reliably under various conditions, including different speakers and environmental backgrounds, enhances system robustness and user experience.

8.2.5 Future Directions

Looking ahead, the field of VSR is poised for significant advancements driven by several key areas of focus. The integration of larger and more diverse datasets will continue to play a pivotal role in improving model generalization and performance. Additionally, the fusion of multimodal data will likely become more sophisticated, leveraging advances in both visual and auditory processing technologies.

Research will also increasingly focus on optimizing models for real-time applications, ensuring that VSR systems can operate efficiently and effectively in practical settings. This includes exploring novel deep learning architectures, such as transformers, which have shown promise in various speech and vision tasks. Furthermore, the development of lightweight models suitable for deployment on mobile and edge devices will expand the accessibility and usability of VSR technology.

In conclusion, this chapter highlights the multifaceted approaches and considerations essential for advancing the field of VSR. By expanding datasets, integrating multimodal data, and continuously optimizing models, researchers and practitioners can significantly enhance the performance, reliability, and applicability of VSR systems. The implications of this research extend beyond academia, offering valuable insights and tools for industrial applications, ultimately shaping the future trajectory of visual speech recognition technology.

8.2.6 Limitations and Future Research Directions

Despite the noteworthy progress made in this study, it is imperative to acknowledge its inherent limitations. The reliance on existing methods for mouth detection, notably the sequential use of Dlib and YOLO, coupled with the specific scope of the dataset, may introduce inherent biases. As a future direction, researchers can delve into exploring innovative approaches to further enhance the accuracy and diversity of visual speech recognition. The study's primary focus on the Arabic language presents an opportunity

for expansion, with potential extensions to encompass a broader array of languages, thus enriching the framework's applicability and impact. Future investigations may also explore advanced model architectures and the integration of additional modalities, such as audio signals, offering a promising avenue to unlock new dimensions in the realm of Visual Speech Recognition research.

Bibliography

[1] T. Afouras, J. Son Chung **and** A. Zisserman. "Deep lip reading: A comparison of models and an online application". **in**(2018): arXiv: 1806.06053. URL: http://arxiv.org/abs/1806.06053.

[2] T. Afouras **andothers**. "Deep audio-visual speech recognition". **in***IEEE Trans. Pattern Anal. Mach. Intell.*: (**december** 2018). Early access. DOI: 10.1109/TPAMI.2018.2889052.

[3] Triantafyllos Afouras, Joon Son Chung **and** Andrew Zisserman. "LRS3-TED: a large-scale dataset for visual speech recognition". **in***arXiv preprint arXiv:1809.00496*: (2018).

[4] Nadia H Alsulami, Amani T Jamal **and** Lamiaa A Elrefaei. "Deep learning-based approach for arabic visual speech recognition". **in***Computers, Materials & Continua*: 71.1 (2022), **pages** 85–108.

[5] Yannis M Assael **andothers**. "Lipnet: End-to-end sentence-level lipreading". **in***arXiv preprint arXiv:1611.01599*: (2016).

[6] Andrew Aubrey **andothers**. "Two novel visual voice activity detectors based on appearance models and retinal filtering". **in***2007 15th European Signal Processing Conference*: IEEE. 2007, **pages** 2409–2413.

[7] Ali Baaloul **andothers**. "An arabic visual speech recognition framework with CNN and vision transformers for lipreading". **in***Multimedia Tools and Applications*: (2024), **pages** 1–35.

[8] Mark Barnard, Eun-Jung Holden **and** Robyn Owens. "Lip tracking using pattern matching snakes". **in***Proc. of the Fifth Asian Conference on Computer Vision*: **volume** 1. 3. 2002.

[9] Y-Lan Boureau **andothers**. "Learning mid-level features for recognition". **in***2010 IEEE computer society conference on computer vision and pattern recognition*: IEEE. 2010, **pages** 2559–2566.

[10] J. S. Chung **and** A. Zisserman. "Learning to lip read words by watching videos". **in***Comput. Vis. Image Understand.*: 173 (**august** 2018), **pages** 76–85.

[11] J. S. Chung **and** A. Zisserman. "Lip reading in profile". **in***Proc. Brit. Mach. Vis. Conf.* 2017, **pages** 1–11.

[12] Joon Son Chung **and** Andrew Zisserman. "Lip reading in the wild". **in***Asian conference on computer vision*: Springer. 2016, **pages** 87–103.

[13] Joon Son Chung **and** Andrew Zisserman. "Out of time: automated lip sync in the wild". **in***Asian conference on computer vision*: Springer. 2016, **pages** 251–263.

[14] Joon Son Chung **andothers**. "Lip reading sentences in the wild". **in***2017 IEEE conference on computer vision and pattern recognition (CVPR)*: IEEE. 2017, **pages** 3444–3453.

[15] Junyoung Chung **andothers**. "Empirical evaluation of gated recurrent neural networks on sequence modeling". **in***arXiv preprint arXiv:1412.3555*: (2014).

[16] Martin Cooke **andothers**. "An audio-visual corpus for speech perception and automatic speech recognition". **in***The Journal of the Acoustical Society of America*: 120.5 (2006), **pages** 2421–2424.

[17] Timothy F Cootes, Gareth J Edwards **and** Christopher J Taylor. "A Comparative Evaluation of Active Appearance Model Algorithms." **in***BMVC*: **volume** 98. 1998, **pages** 680–689.

[18] Timothy F Cootes **and** Christopher J Taylor. "Active shape models—'smart snakes'". **in***BMVC92: Proceedings of the British Machine Vision Conference, organised by the British Machine Vision Association 22–24 September 1992 Leeds*: Springer. 1992, **pages** 266–275.

[19] L. Courtney **and** R. Sreenivas. "Learning from videos with deep convolutional LSTM networks". **in**(2019): arXiv: `1904.04817`. URL: `http://arxiv.org/abs/1904.04817`.

[20] Stephen J Cox **andothers**. "The challenge of multispeaker lip-reading." **in***AVSP*: Citeseer. 2008, **pages** 179–184.

[21] Andrzej Czyzewski **andothers**. "An audio-visual corpus for multimodal automatic speech recognition". **in***Journal of Intelligent Information Systems*: 49.2 (2017), **pages** 167–192.

[22] Navneet Dalal **and** Bill Triggs. "Histograms of oriented gradients for human detection". **in***2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*: **volume** 1. Ieee. 2005, **pages** 886–893.

[23] Patrice Delmas, Pierre-Yves Coulon **and** Vincent Fristot. "Automatic snakes for robust lip boundaries extraction". **in***1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*: **volume** 6. IEEE. 1999, **pages** 3069–3072.

[24] Alexey Dosovitskiy **andothers**. "An image is worth 16x16 words: Transformers for image recognition at scale". **in***arXiv preprint arXiv:2010.11929*: (2020).

[25] Rajneel Dutta **and** Thavavel Vaiyapuri. "LIP READING USING DEEP LEARNING TECHNIQUES". **in**().

[26] Waleed Dweik, Sundus Altorman **and** Safa Ashour. "Read my lips: Artificial intelligence word-level arabic lipreading system". **in***Egyptian Informatics Journal*: 23.4 (2022), **pages** 1–12.

[27] Lamiaa A Elrefaei, Tahani Q Alhassan **and** Shefaa S Omar. "An Arabic visual dataset for visual speech recognition". **in***Procedia Computer Science*: 163 (2019), **pages** 400–409.

[28] Nicolas Eveno, Alice Caplier **and** P-Y Coulon. "A parametric model for realistic lip segmentation". **in***7th International Conference on Control, Automation, Robotics and Vision, 2002. ICARCV 2002.* **volume** 3. IEEE. 2002, **pages** 1426–1431.

[29] S. Fenghour, D. Chen **and** P. Xiao. "Decoder-encoder LSTM for lip reading". **in***Proc. 8th Int. Conf. Softw. Inf. Eng. (ICSIE)*: 2019, **pages** 162–166.

[30] Souheil Fenghour **andothers**. "An Effective Conversion of Visemes to Words for High-Performance Automatic Lipreading". **in***Sensors*: 21.23 (2021), **page** 7890.

[31] I. Fung **and** B. Mak. "End-to-end low-resource lip-reading with maxout CNN and LSTM". **in***Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*: **april** 2018, **pages** 2511–2515.

[32] Amit Garg, Jonathan Noyola **and** Sameep Bagadia. "Lip reading using CNN and LSTM". **in***Technical report, Stanford University, CS231 n project report*: (2016).

[33] Ross Girshick. "Fast r-cnn". **in***Proceedings of the IEEE international conference on computer vision*: 2015, **pages** 1440–1448.

[34] Ross Girshick **andothers**. "Rich feature hierarchies for accurate object detection and semantic segmentation". **in***Proceedings of the IEEE conference on computer vision and pattern recognition*: 2014, **pages** 580–587.

[35] Raymond G Gordon Jr. "Ethnologue, languages of the world". **in***http://www. ethnologue. com/*: (2005).

[36] A. Gutierrez **and** Z. Robert. *Lip reading word classification*. Project Rep. CS231n. Stanford Univ., Stanford, CA, USA, 2017.

[37] S. Nadeem Hashmi **andothers**. "A lip reading model using CNN with batch normalization". **in***Proc. 11th Int. Conf. Contemp. Comput. (IC)*: **august** 2018, **pages** 1–6.

[38] Sarah Hilder, Richard W Harvey **and** Barry-John Theobald. "Comparison of human and machine-based lip-reading." **in***AVSP*: 2009, **pages** 86–89.

[39] Sepp Hochreiter **and** Jürgen Schmidhuber. "Long short-term memory". **in***Neural computation*: 9.8 (1997), **pages** 1735–1780.

[40] D. Hu, X. Li **and** X. Lu. "Temporal multimodal learning in audiovisual speech recognition". **in***Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*: **june** 2016, **pages** 3574–3582.

[41] Sergey Ioffe **and** Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". **in***International conference on machine learning*: PMLR. 2015, **pages** 448–456.

[42] Dawid Jachimski, Andrzej Czyzewski **and** Tomasz Ciszewski. "A comparative study of English viseme recognition methods and algorithms". **in***Multimedia Tools and Applications*: 77.13 (2018), **pages** 16495–16532.

[43] D.-W. Jang **andothers**. "Lip reading using committee networks with two different types of concatenated frame images". **in***IEEE Access*: 7 (2019), **pages** 90125–90131.

[44] Michael Kass, Andrew Witkin **and** Demetri Terzopoulos. "Snakes: Active contour models". **in***International journal of computer vision*: 1.4 (1988), **pages** 321–331.

[45] Katrin Kirchhoff **andothers**. "Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins summer workshop". **in***2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).* **volume** 1. IEEE. 2003, **pages** I–I.

[46] D. Kumar Margam **andothers**. "LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models". **in**(2019): arXiv: `1906.12170`. URL: `http://arxiv.org/abs/1906.12170`.

[47] Yann LeCun **andothers**. “Gradient-based learning applied to document recognition”. **in***Proceedings of the IEEE*: 86.11 (1998), **pages** 2278–2324.

[48] Daehyun Lee, Jongmin Lee **and** Kee-Eung Kim. “Multi-view automatic lip-reading using neural network”. **in***Asian conference on computer vision*: Springer. 2016, **pages** 290–302.

[49] Shu-Hung Leung, Shi-Lin Wang **and** Wing-Hong Lau. “Lip image segmentation using fuzzy clustering incorporating an elliptic shape function”. **in***IEEE transactions on image processing*: 13.1 (2004), **pages** 51–62.

[50] Y. Li **andothers**. “Lip reading using a dynamic feature of lip images and convolutional neural networks”. **in***Proc. IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci. (ICIS)*: **june** 2016, **pages** 1–6.

[51] AW-C Liew, Shu Hung Leung **and** Wing Hong Lau. “Segmentation of color lip images by spatial fuzzy clustering”. **in***IEEE transactions on Fuzzy Systems*: 11.4 (2003), **pages** 542–549.

[52] Shujie Liu **andothers**. “A recursive recurrent neural network for statistical machine translation”. **in***Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*: 2014, **pages** 1491–1500.

[53] Wei Liu **andothers**. “Ssd: Single shot multibox detector”. **in***Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*: Springer. 2016, **pages** 21–37.

[54] Patrick Lucey, Sridha Sridharan **and** David Dean. “Continuous pose-invariant lipreading”. **in***Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech 2008) incorporating the 12th Australasian International Conference on Speech Science and Technology (SST 2008)*: International Speech Communication Association. 2008, **pages** 2679–2682.

[55] Juergen Luettin **and** Stéphane Dupont. “Continuous audio-visual speech recognition”. **in***Computer Vision—ECCV’98: 5th European Conference on Computer Vision Freiburg, Germany, June 2–6, 1998 Proceedings, Volume II 5*: Springer. 1998, **pages** 657–673.

[56] Juergen Luettin, Gerasimos Potamianos **and** Chalapathy Neti. “Asynchronous stream modeling for large vocabulary audio-visual speech recognition”. **in***2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*: **volume** 1. IEEE. 2001, **pages** 169–172.

[57] M. Luo **andothers**. “Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading”. **in***Proc. 15th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*: **march** 2020, **pages** 69–76.

[58] Mohammad H Mahoor **and** Mohamed Abdel-Mottaleb. “Facial features extraction in color images using enhanced active shape model”. **in***7th International Conference on Automatic Face and Gesture Recognition (FGR06)*: IEEE. 2006, 5–pp.

[59] Dilip Kumar Margam **andothers**. “LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models”. **in***arXiv preprint arXiv:1906.12170*: (2019).

[60] Iain Matthews **andothers**. “Extraction of visual features for lipreading”. **in***IEEE Transactions on Pattern Analysis and Machine Intelligence*: 24.2 (2002), **pages** 198–213.

[61] A. Mesbah **andothers**. “Lip reading with Hahn convolutional neural networks”. **in***Image Vis. Comput.*: 88 (**august** 2019), **pages** 76–83.

[62] Kieron Messer **andothers**. "XM2VTSDB: The extended M2VTS database". **in***Second international conference on audio and video-based biometric person authentication*: **volume** 964. Citeseer. 1999, **pages** 965–966.

[63] Stephen Milborrow **and** Fred Nicolls. "Locating facial features with an extended active shape model". **in***Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV 10*: Springer. 2008, **pages** 504–513.

[64] J. Ngiam **andothers**. "Multimodal deep learning". **in***Proc. 28th Int. Conf. Mach. Learn. (ICML)*: 2011, **pages** 689–696.

[65] Quoc Dinh Nguyen **and** Maurice Milgram. "Multi features Active Shape Models for lip contours detection". **in***2008 International Conference on Wavelet Analysis and Pattern Recognition*: **volume** 1. IEEE. 2008, **pages** 172–176.

[66] H. Ninomiya **andothers**. "Integration of deep bottleneck features for audio-visual speech recognition". **in***Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.* 2015, **pages** 1–5.

[67] K. Noda **andothers**. "Audiovisual speech recognition using deep learning". **in***Int. J. Speech Technol.*: 42.4 (**june** 2015), **pages** 722–737.

[68] Kuniaki Noda **andothers**. "Audio-visual speech recognition using deep learning". **in***Applied intelligence*: 42 (2015), **pages** 722–737.

[69] Kuniaki Noda **andothers**. "Lipreading using convolutional neural network". **in***fifteenth annual conference of the international speech communication association*: 2014.

[70] Alfonso Ortega **andothers**. "AV@ CAR: A Spanish Multichannel Multimodal Corpus for In-Vehicle Automatic Audio-Visual Speech Recognition." **in***LREC*: 2004.

[71] Ankur P Parikh **andothers**. "A decomposable attention model for natural language inference". **in***arXiv preprint arXiv:1606.01933*: (2016).

[72] Eric K Patterson **andothers**. "CUAVE: A new audio-visual database for multimodal human-computer interface research". **in***2002 IEEE International conference on acoustics, speech, and signal processing*: **volume** 2. IEEE. 2002, **pages** II–2017.

[73] S. Petridis, Z. Li **and** M. Pantic. "End-to-end visual speech recognition with LSTMS". **in***Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*: **march** 2017, **pages** 2592–2596.

[74] S. Petridis **and** M. Pantic. "Deep complementary bottleneck features for visual speech recognition". **in***Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*: **march** 2016, **pages** 2304–2308.

[75] S. Petridis **andothers**. "Audio-visual speech recognition with a hybrid CTC/attention architecture". **in***Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*: **december** 2018, **pages** 513–520.

[76] S. Petridis **andothers**. "End-to-end audiovisual fusion with LSTMs". **in**(2017): arXiv: 1709.04343. URL: http://arxiv.org/abs/1709.04343.

[77] S. Petridis **andothers**. "End-to-end audiovisual speech recognition". **in***Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*: **april** 2018, **pages** 6548–6552.

[78] S. Petridis **andothers**. "End-to-end multi-view lipreading". **in**(2017): arXiv: 1709.00443. URL: http://arxiv.org/abs/1709.00443.

[79] S. Petridis **andothers**. “Visual-only recognition of normal, whispered and silent speech”. **in***Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*: **april** 2018, **pages** 6219–6223.

[80] M. H. Rahmani **and** F. Almasganj. “Lip-reading via a DNN-HMM hybrid system using a combination of the image-based and model-based features”. **in***Proc. 3rd Int. Conf. Pattern Recognit. Image Anal. (IPRIA)*: **april** 2017, **pages** 195–199.

[81] Joseph Redmon **and** Ali Farhadi. “YOLO9000: better, faster, stronger”. **in***Proceedings of the IEEE conference on computer vision and pattern recognition*: 2017, **pages** 7263–7271.

[82] Joseph Redmon **and** Ali Farhadi. “Yolov3: An incremental improvement”. **in***arXiv preprint arXiv:1804.02767*: (2018).

[83] Joseph Redmon **andothers**. “You only look once: Unified, real-time object detection”. **in***Proceedings of the IEEE conference on computer vision and pattern recognition*: 2016, **pages** 779–788.

[84] Ahmed Rekik, Achraf Ben-Hamadou **and** Walid Mahdi. “A new visual speech recognition approach for RGB-D cameras”. **in***International conference image analysis and recognition*: Springer. 2014, **pages** 21–28.

[85] Shaoqing Ren **andothers**. “Faster r-cnn: Towards real-time object detection with region proposal networks”. **in***Advances in neural information processing systems*: 28 (2015).

[86] R Rohani **andothers**. “Lip processing and modeling based on spatial fuzzy clustering in color images”. **in***International Journal of Fuzzy Systems*: 13.2 (2011), **pages** 65–73.

[87] Leon Rothkrantz. “Lip-reading by surveillance cameras”. **in***2017 Smart City Symposium Prague (SCSP)*: IEEE. 2017, **pages** 1–6.

[88] T. Saitoh **andothers**. “Concatenated frame image based cnn for visual speech recognition”. **in***Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, **pages** 277–289.

[89] Conrad Sanderson. *The vidtimit database*. techreport. IDIAP, 2002.

[90] Amany M Sarhan, Nada M Elshennawy **and** Dina M Ibrahim. “HLR-net: a hybrid lip-reading model based on deep convolutional neural networks”. **in***Computers, Materials & Continua*: 68.2 (2021), **pages** 1531–1549.

[91] Rowan Seymour, Darryl Stewart **and** Ji Ming. “Comparison of image transform-based features for visual speech recognition in clean and corrupted videos”. **in***EURASIP Journal on Image and Video Processing*: 2008 (2007), **pages** 1–9.

[92] B. Shillingford **andothers**. “Large-scale visual speech recognition”. **in**(2018): arXiv: 1807.05162. URL: http://arxiv.org/abs/1807.05162.

[93] Joon Son Chung **andothers**. “Lip reading sentences in the wild”. **in***Proceedings of the IEEE conference on computer vision and pattern recognition*: 2017, **pages** 6447–6456.

[94] Nitish Srivastava **andothers**. “Dropout: a simple way to prevent neural networks from overfitting”. **in***The journal of machine learning research*: 15.1 (2014), **pages** 1929–1958.

[95] T. Stafylakis, M. H. Khan **and** G. Tzimiropoulos. “Pushing the boundaries of audiovisual word recognition using residual networks and LSTMs”. **in***Comput. Vis. Image Understand.*: 176–177 (**november** 2018), **pages** 22–32.

[96] T. Stafylakis **and** G. Tzimiropoulos. “Combining residual networks with LSTMs for lipreading”. **in***Interspeech*: 2017.

[97] T. Stafylakis **and** G. Tzimiropoulos. “Deep word embeddings for visual speech recognition”. **in***Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*: **april** 2018, **pages** 4974–4978.

[98] C. Sui, M. Bennamoun **and** R. Togneri. “Listening with your eyes: Towards a practical visual speech recognition system using deep Boltzmann machines”. **in***Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*: **december** 2015, **pages** 154–162.

[99] C. Sui, R. Togneri **and** M. Bennamoun. “Extracting deep bottleneck features for visual speech recognition”. **in***Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*: **april** 2015, **pages** 1518–1522.

[100] Christian Szegedy **andothers**. “Going deeper with convolutions”. **in***Proceedings of the IEEE conference on computer vision and pattern recognition*: 2015, **pages** 1–9.

[101] Y. Takashima **andothers**. “Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss”. **in***Proc. Interspeech*: **september** 2016, **pages** 277–281.

[102] A. Torfi **andothers**. “3D convolutional neural networks for cross audio-visual matching recognition”. **in***IEEE Access*: 5 (2017), **pages** 22081–22091.

[103] Ashish Vaswani **andothers**. “Attention is all you need”. **in***Advances in neural information processing systems*: 30 (2017).

[104] Paul Viola **and** Michael Jones. “Rapid object detection using a boosted cascade of simple features”. **in***Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*: **volume** 1. Ieee. 2001, **pages** I–I.

[105] Alexander Vorwerk **andothers**. “WAPUSK20-A database for robust audiovisual speech recognition”. **in***Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*: 2010.

[106] M. Wand, J. Koutnik **and** J. Schmidhuber. “Lipreading with long shortterm memory”. **in***Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*: **march** 2016, **pages** 6115–6119.

[107] M. Wand **and** J. Schmidhuber. “Improving speaker-independent lipreading with domain-adversarial training”. **in**(2017): arXiv: `1708.01565 [cs.CV]`. URL: `http://arxiv.org/abs/1708.01565`.

[108] M. Wand, J. Schmidhuber **and** N. T. Vu. “Investigations on end-to-end audiovisual fusion”. **in***Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*: **april** 2018, **pages** 3041–3045.

[109] C. Wang. “Multi-grained spatio-temporal modeling for lip-reading”. **in**(2019): arXiv: `1908.11618`. URL: `http://arxiv.org/abs/1908.11618`.

[110] Shilin Wang **andothers**. “Fuzzy clustering-based approaches in automatic lip segmentation from color images”. **in***Advances in Image and Video Segmentation*: IGI Global, 2006, **pages** 292–317.

[111] X. Weng **and** K. Kitani. “Learning spatio-temporal features with two-stream deep 3D CNNs for lipreading”. **in***Brit. Mach. Vis. Conf.* 2019.

[112] Salah Werda, Walid Mahdi **and** Abdelmajid Ben Hamadou. "Lip localization and viseme classification for visual speech recognition". **in***arXiv preprint arXiv:1301.4558*: (2013).

[113] J. Xiao **andothers**. "Deformation flow based two-stream network for lip reading". **in***Proc. 15th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*: **march** 2020, **pages** 836–842.

[114] Kai Xu **andothers**. "LCANet: End-to-end lipreading with cascaded attention-CTC". **in***2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*: IEEE. 2018, **pages** 548–555.

[115] Chengjia Yang, Xiongfei Li **and** Xiaoli Zhang. "Lip contour extraction of RGB-based improved region growing algorithm". **in***2014 IEEE 5th International Conference on Software Engineering and Service Science*: IEEE. 2014, **pages** 597–600.

[116] S. Yang **andothers**. "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild". **in***Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*: **may** 2019, **pages** 1–8.

[117] Li Yuan **andothers**. "Tokens-to-token vit: Training vision transformers from scratch on imagenet". **in***Proceedings of the IEEE/CVF international conference on computer vision*: 2021, **pages** 558–567.

[118] Marwan Al-Zabibi. "An acoustic-phonetic approach in automatic Arabic speech recognition". phdthesis. Loughborough University, 1990.

[119] X. Zhang **andothers**. "Understanding pictograph with facial features: End-to-end sentence-level lip reading of Chinese". **in***Proc. AAAI Conf. Artif. Intell.* 2019, **pages** 9211–9218.

[120] Y. Zhang **andothers**. "Can we read speech beyond the lips? Rethinking RoI selection for deep visual speech recognition". **in***Proc. 15th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*: **may** 2020, **pages** 851–858.

[121] Guoying Zhao, Mark Barnard **and** Matti Pietikainen. "Lipreading with local spatiotemporal descriptors". **in***IEEE Transactions on Multimedia*: 11.7 (2009), **pages** 1254–1265.

[122] X. Zhao **andothers**. "Mutual information maximization for effective lip reading". **in***Proc. 15th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*: **march** 2020, **pages** 843–850.

[123] P. Zhou **andothers**. "Modality attention for End-to-end audio-visual speech recognition". **in***Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*: **may** 2019, **pages** 6565–6569.

[124] Ziheng Zhou, Guoying Zhao **and** Matti Pietikäinen. "Towards a practical lipreading system". **in***CVPR 2011*: IEEE. 2011, **pages** 137–144.

[125] M. Zimmermann **andothers**. "Visual speech recognition using PCA networks and LSTMs in a tandem GMM-HMM system". **in***Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, **pages** 264–276.