International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

LEXICAL INTERPRETATION OF VISUAL CUES

[1]Ms CH.SUKANYA, [2] A.BHAVANA, [3] CH.SAKETH, [4] G.POOJITHA

[1](Assistant Professor) , CSE. Teegala Krishna Reddy Engineering College Hyderabad.
[2,3,4]B,tech , scholar , CSE. Teegala Krishna Reddy Engineering College Hyderabad.

Abstract

Human lip-reading is a challenging task. It requires not only knowledge of underlying language but also visual clues to predict spoken words. Experts need certain level of experience and understanding of visual expressions learning to decode spoken words, but with the help of deep learning it is possible to translate lip sequences into meaningful words. Computer vision techniques and machine learning algorithms to extract relevant visual features from lip movements. These features are then mapped to corresponding phonemes and linguistic units, enabling the system to recognize and interpret spoken words. This project focuses on the development of a robust system for converting lip movements from video into textual information through the application of Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) algorithms of deep learning. The objective is to bridge the gap between visual cues and linguistic interpretation by leveraging advanced machine learning techniques. By harnessing the power of CNN for feature extraction and GRU for sequential modeling, the system aims to accurately decode and transcribe lip movements into meaningful textual representations. The proposed approach demonstrates a novel integration of computer vision and natural language processing, enabling the conversion of visual data into comprehensible text, thereby enhancing accessibility and communication for individuals with hearing impairments. The experimental results showcase promising accuracy rates, underscoring the potential of this technology to facilitate seamless and efficient communication across diverse linguistic contexts.

KEYWORDS: Lexical Interpretation, Visual Cues, Lip Reading, Deep Learning, Convolutional Neural Network (CNN),Gated Recurrent Unit (GRU), Accessibility, Communication, Machine Learning

1 . INTRODUCTION

The lexical interpretation of visual cues refers to the process of understanding and assigning meaning to various visual elements and stimuli in our environment. This form of interpretation involves the analysis of visual information, such as colours, shapes, textures, and spatial relationships, to derive specific meanings or message. In this context, "lexical" pertains to the use of a vocabulary or set of symbols to represent and convey meaning. When applied to visual cues, the term implies the understanding of how certain visual elements can be interpreted and assigned specific linguistic or conceptual associations. Visual cues play a significant role in communication, perception, and the formation of cognitive representations. They can convey emotions, ideas, and complex information in a concise and powerful manner. The interpretation of these cues often involves a combination of learned associations, cultural influences, and cognitive processes. Understanding the lexical interpretation of visual cues can be vital in various fields, including art, design, marketing, psychology, and communication studies. By grasping the nuances of how different visual elements are interpreted, individuals can effectively convey specific messages, evoke certain emotions, and create engaging and impactful visual experience.

2. LITERATURE SURVEY

Automated lip reading systems initially focused on classifying isolated speech segments in the form of digits and letters [13], [14], [15], [16], [17], and then eventually moved on to longer speech segments in the form of words. The success of automated lip reading was previously constrained by the available training data, as initially, the only audio-visual datasets available were those with isolated speech segments, i.e., digits, alphabet and words [17], [18], [19]. Subsequently every speech segment was treated as a class to recognize.

Title	Methods	Datasets	Accuracy	Limitations
Lip reading using CNN and LSTM, 2017	VGG Net along with SVM	MIRAcL-VC1 dataset	76%	Model trained from scratch did not perform well as size of dataset is small.
Lip Reading Word Classification,2018	VGG16 and attention – based LSTM	MIRACL-VC1 dataset	79%	Does not support larger dataset
Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long ShortTerm Memory,2019	VGG19 and attention – based LSTM	MIRACL-VC1 dataset	88.2%	Requires a good quality of video
Lip Reading Using Convolutional Neural Networks with and without Pre-Trained Models,2020	CNN with AlexNet, GoogleNet	AvLetters [10] dataset	64.4%	Gives more accurate result specifically for alphabet level recognition
Speaker-Independent Speech Recognition using Visual Features,2020	3D CNN	MIRACL-VC1 dataset	76.89%	Limited training data

Table 2.1 Literature Survey

3 . SYSTEM DESIGN

3.1 System Architecture

The system architecture for the lexical interpretation of visual cues involves processing and understanding visual information and mapping it to natural language descriptions or labels. This can be useful in various applications, such as image captioning, object recognition, and visual scene understanding. The lexical interpretation of visual cues involves designing a system that can understand and interpret visual information, converting it into a form that can be processed and understood using language. This is a complex task that often requires a combination of Concurrent Neural Network (CNN) and Gated Recurrent Neural Network (RNN) techniques.

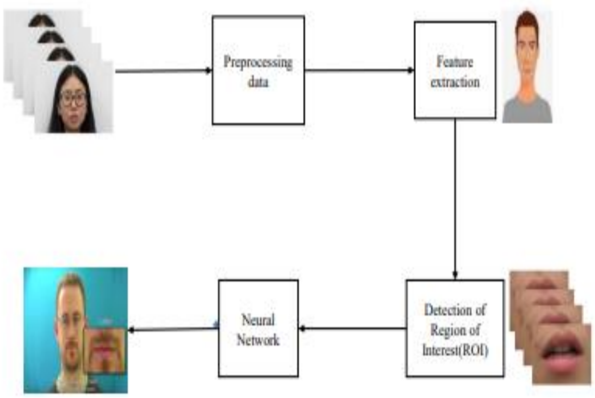

Figure 3.1 Architecture for proposed system

There are components of the architecture:

1. Video Input: Load the video files into your system. Various libraries like OpenCV in Python can help with video loading and processing. This is where visual cues in the form of images or video frames are fed into the system. Various sensors or cameras can be used to capture visual data.

2. Preprocessing data: Techniques like convolutional neural networks (CNNs) to extract visual features from the input video. This step transforms raw pixels into meaningful representations. Preprocessing video data for the lexical interpretation of visual cues involves several steps to extract meaningful information and prepare it for input into your model. Enhance image quality, resize, and normalize to ensure consistent input.

3. Frame Extraction: Extract individual frames from the video. The frame rate can be adjusted based on your requirements. A standard practice is to use a certain number of frames per second (fps) to balance computational cost and temporal information. Extract relevant features from the frames based on lip movements. This might include color histograms, texture features, or deep features from pre-trained convolutional neural networks (CNNs).

4. Detection of Region of interest: For lip reading, detect the region of interest by applying face detection algorithms on video frames. This isolates and extracts the face, focusing on relevant lip movements. Ensure accurate localization and alignment, facilitating precise analysis of lip gestures for robust lexical interpretation in lip reading applications.

5. Neural Network: Neural Network is combination of Conventional Neural Network(CNN) and Gated Recurrent Neural Network(GRR). Convolutional Neural Networks (CNNs) are specialized neural networks designed for image processing and pattern recognition. Comprising layers like convolutional, pooling, and fully connected, CNNs excel at capturing hierarchical features. In the convolutional layer, filters systematically convolve across the input, detecting spatial patterns. Pooling layers reduce dimensionality by down sampling, preserving critical information. These layers, combined with non-linear activation functions, allow CNNs to learn intricate visual hierarchies. Striding and

3.2: Unified Modeling Language (UML):

Unified Modeling Language (UML) is a standardized visual modeling language widely used in software engineering to design and document complex systems. UML provides a set of graphical notations that allow software developers, system architects, and other stakeholders to visualize, specify, construct, and document the artifacts of a software system. It serves as a common language for communication among team members and facilitates the understanding of system architecture and behavior. UML's standardized notation ensures that developers worldwide can interpret and contribute to system design and documentation. UML is not just limited to software development, it can also be applied to various domains like business modeling, system engineering, and process modeling. The flexibility and comprehensiveness of UML make it a valuable tool in the software development life cycle, from initial design and planning to the implementation and maintenance of complex systems.

i. Use Case Diagram A use case diagram is a type of Unified Modeling Language (UML) diagram that illustrates how a system interacts with its users or external entities. It depicts the various use cases, actors, and their relationships within a system. This is a diagram or set of diagrams that together with additional documentation shows what the proposed system is designed to do. A use case diagram is used to represent the dynamic behavior of a system. It encapsulates the system's functionality by incorporating use cases, actors, and their relationships. It models the tasks, services, and functions required by a system/subsystem of an application. It depicts the high-level functionality of a system and also tells how the user handles a system.

Actor:

User: Represents the individuals interacting with the system, users might include individuals trying to understand , upload and gets the results.

Use Cases:

It represents the primary functionality of the system. It shows the flow of getting output from the uploaded video. This could include recognizing and understanding words, symbols, or textual elements within a visual context.

Relationships:

Connects an actor to a use case, indicating that the actor is involved in the described functionality

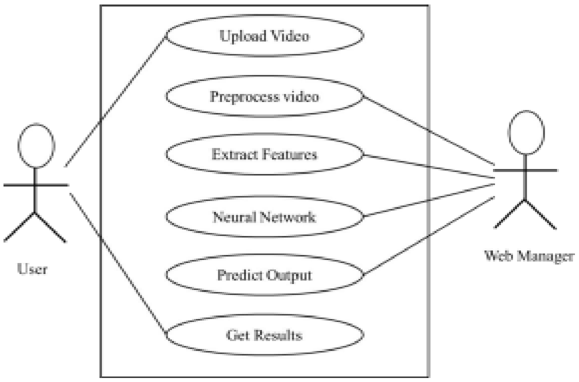

Figure 3.2 Usecase Diagram for Proposed System

ii. Activity Diagram

Activity diagram is another important diagram in UML which describes the dynamic aspect of the proposed system. It is basically a flowchart to represent the flow from one activity to another activity. The activities can be described as an operation of the proposed system. Activity diagram gives a high level understanding of the systems functionalities. Before drawing the activity diagram, we must have a clear understanding about the elements to use. In the proposed system,

the main elements of an activity are the activity itself. An activity is a function performed by the proposed system.

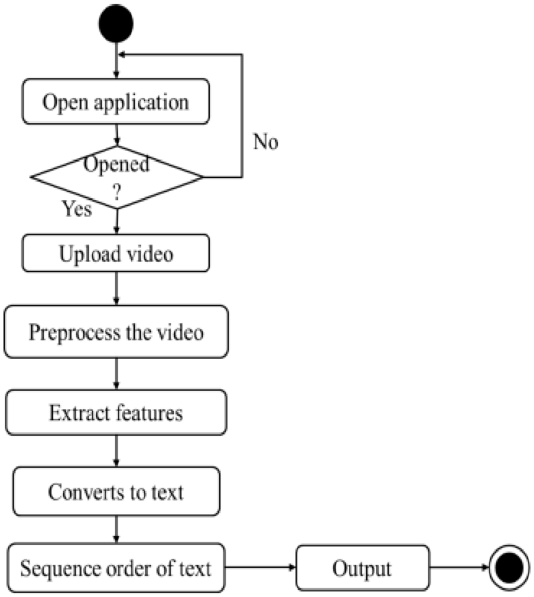

Figure 3.3 Activity Diagram for proposed System

4. CONCLUSION

A lip reading system, based on neural networks, has been created to anticipate sentences spoken in silent videos, encompassing a broad vocabulary range. This lexicon-free system relies solely on visual cues derived from a limited set of distinct lip movements, known as visemes, and exhibits resilience to varying lighting conditions. Validation on the grid dataset revealed a substantial enhancement in word classification accuracy compared to current state-of-the-art methods.Ongoing research aims to identify a more suitable neural network architecture to enhance the system's generalization capabilities, specifically targeting a higher ratio of training to test samples. Additionally, the efficient conversion of visemes to words is identified as a critical aspect when employing visemes as a classification scheme for lip reading sentences. Despite the proposed system achieving high accuracy in viseme classification, a notable drop in word classification accuracy was observed post-conversion in experiments. Hence, exploring alternative approaches for this conversion process is imperative. For perplexity analysis-based conversion, it is crucial to consider various global optimization methods while minimizing computational overhead.

5. FUTURE ENHANCEMENT

Develop advanced multimodal models that can effectively integrate visual and textual information. These models should be able to understand the nuances of both modalities and create a more comprehensive interpretation. Implement pretraining strategies that involve joint training on large datasets containing both visual and textual information. This can help the model learn cross-modal representations and improve its ability to connect visual cues with linguistic expressions. Improve the model's ability to recognize and interpret fine-grained visual details. This could involve advancements in object recognition, segmentation, and understanding complex visual scenes. Develop techniques for semantic segmentation that go beyond basic object recognition to understand the relationships and interactions between objects in an image. This can provide a richer context for lexical interpretation. Enhance models to consider broader contextual information when interpreting visual cues. Understanding the context in which visual information is presented can significantly improve the accuracy and relevance of lexical interpretation. Develop models that can incrementally update their knowledge as they encounter new visual concepts. This is important for keeping the model up-todate with evolving visual cues and understanding emerging trends. Integrate explainability features into the model to provide insights into how it arrived at a particular interpretation. This is crucial for building trust in the model's capabilities and understanding potential biases in its interpretations. Customize models for specific domains to improve their accuracy in interpreting visual cues within those domains. This could involve fine-tuning on domain-specific datasets or incorporating domain-specific knowledge into the models. Improve the speed and efficiency of lexical interpretation in real-time scenarios. This is particularly important for applications such as augmented reality, where quick and accurate interpretation of visual cues is essential.

6. REFERENCES

[1] A. Fernandez-Lopez and F. M. Sukno, ‘‘Survey on automatic lip-reading in the era of deep learning,” Image Vis. Comput., vol. 78, pp. 53–72, Oct. 2018.

[2] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, ‘‘A review of recent advances in visual speech decoding,’’ Image Vis. Comput., vol. 32, no. 9, pp. 590–605, Sep. 2014.

[3] T. Afouras, J. S. Chung, and A. Zisserman, ‘‘Deep lip reading: A comparison of models and an online application,’’ in Proc. Inter speech, Sep. 2018, pp. 3514–3518.

[4] A. B. Mattos, D. Oliveira, and E. Morais, ‘‘Improving viseme recognition using GAN-based frontal view mapping,’’ in Proc. Analysis and Modeling of Faces and Gestures (CVPR), Jun. 2018, pp. 2148–2155.

[5] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, ‘‘Improving language understanding by generative pre-training,’’ OpenAI, San Francisco, CA, USA, Tech. Rep., 2018.

[6] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, ‘‘End-to-end audiovisual speech recognition,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 6548–6552.

[7] S. Petridis, J. Shen, D. Cetin, and M. Pantic, ‘‘Visual-only recognition of normal, whispered and silent speech,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 6219–6223.

[8] M. Wand, J. Schmidhuber, and N. T. Vu, ‘‘Investigations on end- toend audiovisual fusion,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 3041–3045.

[9] K. Xu, D. Li, N. Cassimatis, and X. Wang, ‘‘LCANet: End-to-end lipreading with cascaded attention-CTC,’’ in Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), May 2018, pp. 548–555.

[10] T. Afouras, J. S. Chung, and A. Zisserman, ‘‘LRS3-TED: A large-scale dataset for visual speech recognition,’’ 2018, arXiv:1809.00496. [Online]. Available: https://arxiv.org/abs/1809.00496.