International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.039 ISSN: 2581-4621

LIP MOVEMENT DETECTION USING 3D CONVOLUTION AND RESNET

[1]Kothapeta Sai Shree, [2]Ameti Ganga Prashanth, [3]Madduri Sneha Sri Vardhini, [4]Bhimavarapu Sai Maneesh reddy,[5]Yerramshetty Venkata Nithin
[1]Student, [2]Student, [3]Student, [4]Student, [5]Student
[1,3,4,5]Computer Science & Engineering-Artificial Intelligence and machine Learning, Hyderabad Institute of Technology & Management, Hyderabad, India
[2]Bachelor of Computer Science, Girraj Government College, Nizamabad

Abstract: Lip reading is crucial for communication, especially in noisy environments or situations lacking audio, aiding those with hearing impairments. However, accurately recognizing lip movements is challenging due to variations among individuals and the rapid, subtle nature of lip motions during speech. Deep learning techniques like RESNET and 3D CNN show promise in addressing these challenges. RESNET extracts spatial lip information, while 3D CNN models temporal dynamics, enhancing accurate lip movement interpretation. Training involves initializing the model with pre-trained RESNET weights from ImageNet and fine-tuning it on lip movement datasets, using appropriate loss functions. Evaluation metrics like accuracy, precision, recall, or F1 score assess the model's performance. Once trained, the model can analyse lip movements in real-time or predict spoken words from lip video frames. This RESNET and 3D CNN fusion facilitates robust lip movement detection, benefiting lip reading and speech recognition applications. By capturing spatial and temporal aspects of lip motions, it improves accuracy, offering opportunities to enhance communication accessibility for those with hearing impairments. Continued advancements in deep learning promise to further refine lip reading systems, contributing to more reliable communication aids.
Keywords: 3D densely connected CNN; Resnet; Activation Function; Feature extraction; Data preprocessing, Visual speech recognition.

1.Introduction

In the realm of deep learning, the advent of technologies like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) has paved the way for breakthroughs in various applications, one of the most fascinating being lip reading. Lip reading, which interprets lip, tongue, and face movements to understand speech, enhances communication accessibility for the hard-of-hearing and improves security systems. This article explores leveraging 3D convolutional neural networks for lip reading, offering a comprehensive guide from environment setup to model evaluation. By combining CNNs with robust language models, we transcend traditional speech recognition barriers. We provide insights on building, training, and evaluating a 3D convolution model, highlighting its future implications for lip reading and communication technology.

In 1976, McGurk and McDonald demonstrated that speech perception is influenced by vision, which is called the McGurk effect.[1] This effect indicates the necessity of matching both auditory and visual phonemes to perceive pronounced phonemes correctly. Vision plays a crucial role in speech understanding, and the importance of utilizing visual information to improve the performance and robustness of speech recognition has been demonstrated.[2][3] Although acoustic information is richer than visual information when speaking, most people rely on watching lip movements to fully understand speech.[2]
However, comprehending oral language using visual information alone, especially in the absence of context, can be challenging because it is difficult to understand lipreading actuations such as lip,

tongue, and teeth movements without context. [4] The development of a visual speech recognition (VSR) system has enormous potential for various practical applications, such as speech recognition in noisy environments, biometric identification for security, communication in underwater environments, and silent movie analysis, and it can positively affect patients with speech impairments.[5][6] Recently, several researchers have investigated the possibility of developing a VSR system by decoding speech using only visual information to mimic human lipreading capability [7][8].A major challenge is the distinction of words with similar pronunciation, called homophones [9].

For example, although some words, such as pack, back, and mac, differ in their sound, the characters (e.g., [p], [b], and [m]) produce almost identical lip movements, thereby making them difficult to distinguish. Another technical limitation of traditional VSR systems is that visual information does not provide sufficient data for learning words such as “a”, “an”, “eight”, and “bin” because their length is no longer than 0.02 s [10]. The process of lip movement detection starts with the collection or creation of a static video dataset containing lip movement sequences.

Each video is then divided into frames, and corresponding labels or annotations are assigned to indicate the spoken words or lip movement classes. The availability of a well-annotated dataset is crucial for training the models effectively. Next, the chosen network architecture, either ResNet or 3DCNN, is employed to process the video frames. When using 3DCNN, the model is trained to extract high-level visual features from each individual frame. These features capture important patterns and representations related to lip movements, which are essential for accurate analysis and recognition. In our experiment, we compared the accuracy and efficiency of our architecture with those of other visual feature extraction models with excellent performance on a benchmark dataset.[11] Automatic lipreading plays an important role in communications in noisy environments, e.g. in stadiums and bars where noises overcome speaking signals. The lipreading is traditionally viewed as a supplement to speech recognition.[12].

Lipreading faces three main challenges: collecting large, diverse datasets due to variations in lip shapes and speaking styles; ensuring real-time processing for effective communication; and handling illumination and texture variations in videos. To address these, we propose a novel framework that uses alignment of random forest manifolds for lipreading. This approach maps video clips to a low-dimensional manifold of motion patterns, using patch trajectories. An unsupervised density random forest estimates trajectory affinities, providing fast training and online testing with good generalization. Our experiments combine depth and colour videos to mitigate illumination and texture issues.

2.Related Work

In recent years, there has been a quantum leap in the performance of visual speech recognition systems, thanks to the advances in deep learning techniques [13, 14,15,16] and the availability of large-scale datasets.[17,18] An example based on this approach is LipNet [19], which uses a spatio-temporal frontend, with 3D and 2D convolutions for generating the features, followed by two layers of BLSTM. A deeper architecture than LipNet [19] is used by [20], who propose a residual network with 3D convolutions to extract more powerful representations.

Lip movements are inherently dynamic, with significant changes occurring over short periods. By adopting 3D convolution, it is possible to capture these temporal patterns and dynamics effectively. This capability is crucial for lip detection applications where understanding the sequence of lip movements is essential for accurate interpretation. The Let's Talk! project leverages this technology using a series of 3D CNNs to extract spatiotemporal characteristics from a sequence of fifteen RGB 160×80 images sourced from the MobLip dataset. The architecture employs various parameters such as padding sizes, strides, and kernel sizes to optimize feature extraction from these dynamic visual inputs.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.039 ISSN: 2581-4621

Traditional lip detection methods, such as edge detection or template matching, primarily rely on static image features and often fall short when dealing with the complex, dynamic nature of lip movements. These methods do not account for the temporal aspects that are critical for understanding speech and expressions accurately. In contrast, 3D convolution-based approaches can learn the spatio-temporal features of lip movements, providing a more robust and accurate system. This method proves to be superior, especially in challenging scenarios involving variable lighting conditions, different head poses, or partial occlusions, making it a more reliable option for real-world applications.

3. Architecture

3D Convolution Neural Network

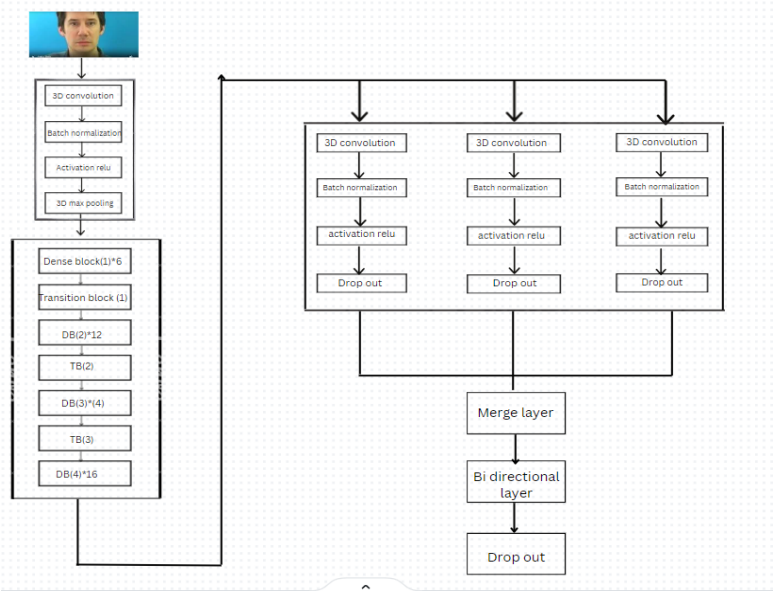

Fig: Architecture of 3D CNN for lip movement detection

The proposed lip movement detection system utilizes a combination of 3D Convolutional Neural Networks (3D CNNs), Long Short-Term Memory (LSTM) networks, Max pooling layers, and activation functions to effectively analyze and classify lip movements in videos. The process begins with 3D CNNs, which extract spatiotemporal features from the video frames by applying 3D filters that consider both spatial and temporal dimensions. This allows the model to analyze local patterns and detect relevant features from the input video frames.

Following the 3D convolutional layers, Max pooling layers are applied to down sample the spatial dimensions of the feature maps, thus reducing computational complexity while retaining the most salient features. Specifically, the proposed system includes 38 Max pooling layers after certain Conv3D layers to optimize the spatial dimensions. Activation layers, such as ReLU (Rectified Linear Unit), are introduced after Conv3D layers to incorporate non-linearity into the network, enabling it to capture complex relationships between inputs and outputs. The Time Distributed layer then converts the 3D output of the previous layers into a 2D tensor, preparing it for the recurrent layers. Subsequently, a bidirectional LSTM layer with 128 units processes the input sequence in both forward and backward directions, capturing the temporal dynamics of lip movements. By combining these components, the system effectively extracts, processes, and analyzes lip movements, making it suitable for applications like lip reading.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.039 ISSN: 2581-4621

RESNET

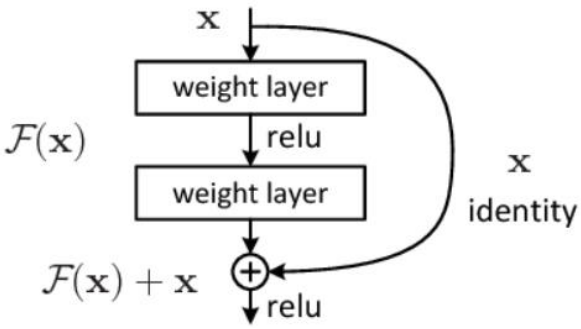

Fig: Resnet for Lip Movement Detection

ResNet (Residual Neural Network) has shown great effectiveness in various computer vision tasks, including lip movement detection, due to its ability to mitigate the vanishing gradient problem in deep networks. ResNet introduces residual connections that allow for more efficient gradient propagation during training, enabling the successful training of much deeper networks. This is particularly beneficial for lip movement detection, which requires capturing subtle and discriminative visual features from the lip region. ResNet's deep architecture and its ability to learn complex representations enable it to extract high-level features from lip images, aiding in distinguishing between different lip movements or phonemes and improving overall detection accuracy. The model's capacity to generalize well to unseen lip movement data is enhanced by these learned hierarchical features, which effectively capture important visual cues related to lip movements, even under varying lighting conditions or across different individuals.
In the implementation, a pretrained ResNet50 model, initialized with weights from the ImageNet dataset, is used. A Sequential model is created to stack layers in sequence. Initially, a Conv2D layer with 3 filters and a kernel size of (1, 1) is applied, with padding to maintain spatial dimensions. The input shape is specified as (75, 46, 140, 1), representing the temporal dimension (number of frames), width, height, and channels, respectively. A Time Distributed layer applies the same ResNet50 layers to each frame independently, followed by a Time Distributed Flatten layer to convert the output into a 1D vector. A bidirectional LSTM layer with 128 units processes the input sequence in both forward and backward directions, capturing temporal dependencies. The softmax function at the output layer normalizes the predictions into a probability distribution, assigning high probabilities to the correct lip movement class. During inference, the class with the highest probability is considered the predicted lip movement.

4. Experiments and Results

This section describes the used dataset, data pre-processing, data augmentation, and implementation.

4.1. Dataset

The dataset consists of 1000 diverse static videos, each capturing individuals speaking or moving their lips, with an average duration of 3 seconds. These videos vary in length to include a wide range of lip movements and expressions. Additionally, the dataset features different lighting conditions, such as varying brightness levels and angles, to ensure robustness in real-world scenarios. This diversity in video duration, speakers, and lighting conditions helps train and evaluate models to accurately detect and recognize lip movements, enhancing their performance in various applications.

4.2 Data Pre-processing

Frames are extracted from the videos using the OpenCV library, which allows for the extraction of individual frames from a video stream. Each frame captures a specific moment in time as a single image. Preprocessing steps include resizing the images to a specific resolution, normalizing pixel

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.039 ISSN: 2581-4621

values, and augmenting the data with techniques like rotation, flipping, or adding noise. Additionally, the dataset is split into training and testing subsets to accurately evaluate the model's performance.

4.3 Data Augmentation

To enhance dataset diversity and improve model robustness, data augmentation techniques were applied. These include random cropping, flipping, rotation, and adjustments in brightness or contrast. Data augmentation increases the dataset size and introduces variations in lip movements, speaker positions, and lighting conditions. The dataset is carefully labeled, split into subsets, augmented, and preprocessed to provide a comprehensive and representative collection of static videos for training and evaluating lip movement detection models.

4.4 Implementation

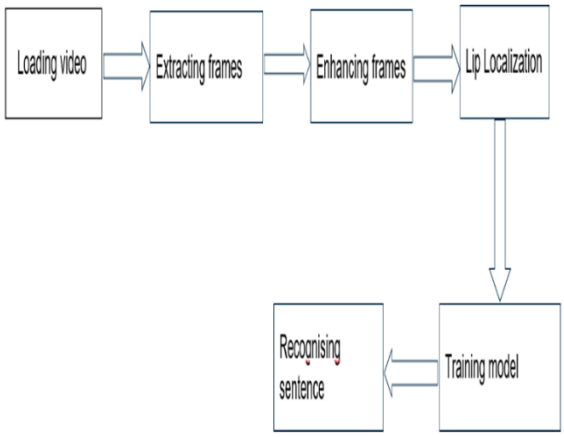

Fig: Design of our Lip Movement Detection model

Here's a clear and concise explanation of the process for preparing video data for lip movement detection:

1. Loading Video: A video processing library is used to load the static video file. This library provides functions to read the video and access its frames, with each frame represented as an array of pixels.
2. Extracting Frames: After loading the video, individual frames are extracted. This involves iterating through each frame and saving it as an individual image. These frames can be stored as separate image files or kept in memory for further processing and will serve as the input data for the lip movement detection model.
3. Enhancing Frames: Enhancing the frames improves their quality before further processing. Enhancements may include adjusting brightness and contrast or applying image denoising techniques to reduce noise. These enhancements help in obtaining clearer and more accurate information about the lip region.
4. Lip Localization: Lip localization identifies the mouth area (Region of Interest, or ROI) in each video frame. This can be done using techniques such as lip shape detection, skin color segmentation, or deep learning methods. Accurate localization is crucial for analyzing lip movements effectively.

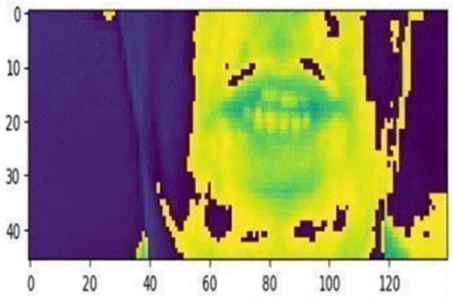

Fig: Lip Localization

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.039 ISSN: 2581-4621

5. Training Using 3D CNN and ResNet: After extracting the frames and localizing the lips, the next step is to train the lip movement detection model. This involves using a 3D Convolutional Neural Network (3D CNN) or a modified Residual Network (ResNet) architecture. The training process uses the extracted frames and their corresponding lip movement labels. The model learns to identify patterns and features in the lip region over time, which allows it to accurately predict various lip movements. During training, optimization algorithms such as backpropagation and gradient descent are employed to adjust the model's parameters. This adjustment process aims to minimize the difference between the predicted lip movements and the actual ground truth labels, improving the model's accuracy.

5. Discussion and Conclusions

After training, the lip movement detection model can predict lip movements in new static videos. Each frame of the video is processed through the trained model, which analyzes the lip movements and generates predictions. These predictions can be converted into text, where the detected lip movements are represented as textual labels or transcriptions. This text-based output provides detailed information about the lip movements in the video and can be further utilized for tasks like speech recognition or emotion recognition. The following table provides a detailed hyperparameters further describing the proposed end-to-end lip-reading architecture.

Layer (type)	Output Shape	Parameters
conv3d (Conv3D)	(None, 75, 46, 140, 128)	3584
activation (Activation)	(None, 75, 46, 140, 128)	0
max_pooling3d (MaxPooling3D)	(None, 75, 23, 70, 128)	0
conv3d_1 (Conv3D)	(None, 75, 23, 70, 256)	884992
activation_1 (Activation)	(None, 75, 23, 70, 256)	0
max_pooling3d_1 (MaxPooling 3D)	(None, 75, 11, 35, 256)	0
conv3d_2 (Conv3D)	(None, 75, 11, 35, 75)	518475
activation_2 (Activation)	(None, 75, 11, 35, 75)	0
max_pooling3d_2 (MaxPooling 3D)	(None, 75, 5, 17, 75)	0
time_distributed (TimeDistributed)	(None, 75, 6375)	0
bidirectional (Bidirectional)	(None, 75, 256)	6660096
dropout (Dropout)	(None, 75, 256)	0
bidirectional_1 (Bidirectional)	(None, 75, 256)	394240
dropout_1 (Dropout)	(None, 75, 256)	0
dense (Dense)	(None, 75, 41)	10537

Total params: 8,471,924
Trainable params: 8,471,924
Non-trainable params: 0

This guide explores the use of 3D convolution for lip detection, emphasizing its advantages over traditional methods in communication technology and accessibility. By leveraging advanced

architectures and dynamic spatiotemporal features, these models achieve high accuracy in interpreting complex lip movements, demonstrating the future potential of speech recognition technologies. Readers gain a solid understanding of the principles and applications of 3D convolution, observing its effectiveness in various challenging scenarios. Beyond technical insights, the guide offers a roadmap for deep learning enthusiasts in lip reading, highlighting future research directions and the importance of improving model accuracy to develop superior communication solutions using 3D convolution.

Author Contributions: Conceptualization-Vardhini; methodology-Prashanth and Sai Shree; software- Nithin and Maneesh; validation-Prashanth; formal analysis and investigation-Vardhini; data curation-Nithin; writing(review and editing) - Maneesh and Sai Shree; supervision,

References

[1] Harry McGurk and John MacDonald. Hearing lips and seeing voices. Nature, 264:746–748, 1976.
[2] AG Chitu and LJM Rothkrantz. Automatic visual speech recognition, pages 95–120. Intech, 2012. Hoofdstuk 6.
[3] Junfei Qiao, Gongming Wang, Wenjing Li, and Min Chen. An adaptive deep qlearning strategy for handwritten digit recognition. Neural Networks, 107:61–71, 02 2018.
[4] Randolph D. Easton and Marylu Basala. Perceptual dominance during lipreading. Perception & Psychophysics, 32:562–570, 1982.
[5] Dimitris Kastaniotis, Dimitrios Tsourounis, and Spiros Fotopoulos. Lip reading modeling with temporal convolutional networks for medical support applications. 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pages 366–371, 2020.
[6] Haoran Zhao, Bowen Zhang, and Zhanhang Yin. Lip-corrector: Application of bertbased model in sentence-level lipreading. Journal of Physics: Conference Series, 1871, 2021.
[7] Adriana Fernandez-Lopez and Federico M. Sukno. Survey on automatic lip-reading in the era of deep learning. Image Vis. Comput., 78:53–72, 2018.
[8] ahmed rekik, Achraf Ben-Hamadou, and Walid Mahdi. Human machine interaction via visual speech spotting. 10 2015.
[9] Mingfeng Hao, Mutallip Mamut, Nurbiya Yadikar, Alimjan Aysa, and Kurban Ubul. A survey of research on lipreading technology. IEEE Access, 8:204518–204544, 2020.
[10] Jingyun Xiao, Shuang Yang, Yuanhang Zhang, S. Shan, and Xilin Chen. 3d feature pyramid attention module for robust visual speech recognition. ArXiv, abs/1810.06178, 2018.
[11] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition (l). The Journal of the Acoustical Society of America, 120:2421–4, 12 2006.
[12] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W. Senior. Recent advances in the automatic recognition of audiovisual speech. Proc. IEEE, 91:1306–1326, 2003.
[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in NIPS, 2012. 1
[14] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in International Conference on Learning Representations, 2015. 1
[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” arXiv preprint arXiv:1512.03385, 2015. 1, 6
[16] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. ICML, 2015. 1
[17] J. S. Chung and A. Zisserman, “Lip reading in the wild,” in Proc. ACCV, 2016. 1
[18] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” in Proc. CVPR, 2017. 1, 3

DOI:10.46647/ijetms.2024.v08i03.039 ISSN: 2581-4621

[19] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: Sentence-level lipreading," arXiv preprint arXiv:1611.01599, 2016. 1, 2, 3
[20] J. S. Chung and A. Zisserman, "Lip reading in profile," in Proc. BMVC., 2017. 1, 3
[21] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with lstms for lipreading," in Interspeech, 2017. 1, 2, 3, 6