iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 Issue: VI Month of publication: June 2023

DOI: https://doi.org/10.22214/ijraset.2023.54180

Lip Movement Detection Using 3D Convolution and Resnet

O Obulesu[1], Teneti.Sanjana[2], V Rupa Sree[3], Saahithya D V[4], B Srija Reddy[5]
[1, 2, 3, 4, 5]Department of CSE, G. Narayanamma Institute of Technology & Science, Hyderabad, TS, India

***Abstract:** Recognition of Lip movements has become one of the most challenging tasks and has crucial applications in the contemporary scenario. Being able to see speech helps people communicate better, especially in challenging listening environments like when there is a background noise and video surveillance when there is no audio. Lip reading is a technique primarily used by deaf people or those who have some form of hearing impairment. It's a way of understanding and communicating with others who might not be familiar with another form of inclusive communication, such as sign language. Lip-reading technology mainly includes face detection, lip localization, feature extraction, training the classifier and finally recognising the word or sentence through lip movement. Many developments have taken place in this growing field using various deep learning-based techniques. An intelligent system will be trained by giving users lip-movement frames sequences as input and will identify lip movement and the said word using 3D convolution and ResNet . This project does analysis over various deep learning models and other datasets. This study also aims to find out the optimal architecture suitable for building a new model with high accuracy for lip movement detection.*

***Keywords:** Lip Movement Detection, Deep Learning,3D Convolution, Resnet, Lip-reading*

I. INTRODUCTION

Lip movement detection is a technique that aims to interpret and understand spoken language by analysing visual information from the speaker's face, with a specific focus on lip movements and facial cues. It involves capturing and analysing the movements and gestures made by the lips during speech to extract valuable linguistic information. Theprocess of lip movement detection involves using computer vision algorithms to track and analyse the shape, position, and motion of the lips in real-time. By capturing and interpreting these visual cues, machinescan gain insights into the spoken language, helping totranscribe, recognize, or understand the words or sentences being spoken. Lip movement detection is particularly useful in scenarios where the audio quality is poor, or in situations where audio alone may not be sufficient to accurately interpret speech. By leveraging the visual information from the speaker's face, machines can enhance speech recognition systems andimprove the accuracy of transcriptions, especially in noisy environments or when there are speech impediments. It has gained significant importance across a range of fields, encompassing speech recognition, lip reading, emotion analysis, and human- computer interaction. By comprehending and interpreting lip movements, machines can extract valuable information about spoken language, the expression of emotions through facial cues, and facilitate more seamless communication between humans and computers. In the past, lip movement detection predominantly relied on manual observationand analysis, which was subjective and demanded considerable time and effort. However, the emergenceof computer vision and machine learning techniques has revolutionized this domain by introducing automated methods that improve accuracy and efficiency. The integration of computer vision and machine learning techniques has streamlined and improved the process of lip movement detection.

By detecting the movement of lips from video input and generating corresponding text output is a fascinating area of research in computer vision andnatural language processing. This task involves analysing the visual information of lip motion in the video frames and converting it into meaningful textualrepresentations. To accomplish this, severaltechniques have been developed that leverage the power of deep learning algorithms. One popular approach is to use Convolutional Neural Networks (CNNs) to extract spatial features from individual frames of the video. CNN's are capable of learning complex patterns and structures in images, making them well-suited for lip feature extraction. In the context of lip movement detection, a common practiceis to divide the video into smaller temporal segments or sequences.

These sequences capture the temporal dynamics of lip motion, allowing the model to understand the progression and changes in lip shapes over time. This is where 3D CNN comes into play. Byextending the traditional 2D convolutions to the temporal dimension, 3D CNNs enable the extraction of spatiotemporal features from the video sequences, providing a richer representation of lip movement.

The movement of lip detection has seen remarkable progress through the integration of deep learning algorithms, particularly with the use of advanced architectures like RESNET (Residual Neural Network) and 3D CN (Convolutional Neural Network). RESNET is known for its ability to train deep neural networks by addressing the problem of vanishing gradients, allowing for more effective feature extraction and representation. On the other hand, 3D CNN extends the conventional 2D convolution operation to the temporal dimension, making it ideal for capturing the temporal dynamics oflip movement. The combination of RESNET and 3D CNN has revolutionized the field of lip detection by enhancing the accuracy and robustness of the models.These techniques have proven effective in extracting high-level spatiotemporal features from lip sequences,enabling the models to capture subtle variations in lip movement. By incorporating both spatial and temporal information, the RESNET-3D CNN models can better discriminate between different lip shapes and movements, even in challenging conditions such as variations in lighting, pose, and occlusions. The movement of lip detection using RESNET and 3D CNN holds great potential in various applications. Forinstance, in the field of assistive technology, it can aidindividuals with speech impairments by converting their lip movements into understandable speech. Moreover, it can contribute to the development of more accurate lip reading systems, which can be used for improving communication in noisy environments or for enhancing speech recognition systems.

II. RELATED WORKS

This work analyses various approaches and methods used in lip movement recognition. Initially, techniquessuch as Hidden Markov Model (HMM), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Gaussian Mixture Modeling (GMM), and Random Forest were employed to extract and classify visual features. These methods were used by different researchers, including Rekik et al., who utilized HMM and depth data to perform lip reading. Gergen et al., who applied LDA within a HMM and GMM framework using the GRID dataset. In recent times, deep learning approaches have gained popularity due to improved access to Graphic Processing Units (GPUs) and have shown promising results in lip reading.

Ngiam et al. used Restricted Boltzmann machinesfor sound and video analysis, while LipNet, proposedby Assael et al., introduced an end-to-end sentence- level modelling approach using recurrent neural networks. Gutierrez et al. experimented with acombination of CNN and LSTM for lip image sequence training. Chung et al. presented the Watch, Listen, Attend, Spell (WLAS) method, which focuseson understanding lip movements and speechinterpretation, along with a training scheme for faster learning. Afouras et al. used 3D-CNN, ResNet, and classifiers like Bidirectional LSTM (Bi-LSTM) and Language Model as feature extractors. Wand et al. combined a Feed-forward network and LSTM layer, while Shillingford et al. utilized 3D-CNN and Bi-LSTM with a Finite-state transducer for sentence classification. Wang proposed a 3D-CNN with deep layered Bidirectional Convolutional LSTM (Bi-Conv-LSTM) for word classification. Petridis et al. employed 3D-CNN, ResNet, and Bi-GRU as a classifier, and other techniques such as Depthwise CNN and Attention Encoder combined with LanguageModel were also used for lip movement recognition.

III. PROPOSED METHODOLOGY

The main objective of this project is to develop a deep learning model capable of accurately recognizing spoken sentences by analysing the lip movementsobserved in input videos. The proposed approach involves a series of essential steps. Firstly, the lip videos undergo a pre-processing stage to enhance their suitability for subsequent analysis. This encompasses various tasks, such as converting the videos into an appropriate format, resizing or cropping frames, and applying techniques to improve the quality of the lip region. Next, the videos are processed to extract keyframes, which are frames that capture important information or significant changes within the video sequence. This step serves to reduce computational complexity and focus on relevant frames relevant to lip-reading. Once the keyframes are obtained, the precise localization of the mouth region becomes a critical task for effective lip-reading. Computer visiontechniques, are utilized to accurately isolate and extract the mouth region from the keyframes.

After localizing the mouth regions, features are extracted from the data using a Convolutional Neural Network (CNN). CNNs are highly effective models for learning discriminative features from images. By inputting the localized mouth images intothe CNN, the model can learn and capture high-level representations that encapsulate the unique lip movements. To capture the temporal dependencies, present in the lip movements, the proposed approach employs a Long Short-Term Memory (LSTM) model.LSTMs are a specific type of recurrent neural networkknown for their ability to model sequential information effectively. By incorporating LSTMlayers into the model, it becomes possible to learn patterns in the lip movements over time and account for the sequential nature of spoken sentences. Finally,the lip-reading results are predicted using a SoftMax layer. The SoftMax function normalizes the output ofthe fully connected layers and assigns probabilities toeach possible class. This ensures that the sum ofprobabilities equals one. The predicted sentence is determined by selecting the class with the highest probability, indicating the recognized sentence.

A. Loading Video And Extracting Frames

Load the video data containing lip movement sequences. Various libraries such as OpenCV or imagecan be used to read the video frames. The load video function takes a path as an argument, which representsthe path to a video file. The function returns a list of normalized frames from the video. The function startsby creating a Video Capture object named by passingthe path argument to it. This object is used to read frames from the video file. The frames extracted fromthe video are stored in an empty list named frames Then the number of frames in the video are calculatedand each frame is read from the video. The read method is used on video capture object to read the nextframe. The return value indicates whether the frame was successfully read, and frame contains the actual frame data.

Fig.1: Frames extracted from video

B. Enhancing Frames

Now, the necessary pre-processing techniques are applied to enhance the quality and relevance of the frames. Pre-processing techniques like grayscaleconversion is used

The frames are converted from RGB to grayscale usingTensorFlow's tf.image.rgb_to_grayscale function.This operation converts the frame from a 3-channel (RGB) image to a single-channel (grayscale) image. The grayscale image represents the intensity values of the original image without considering colour information. Grayscale images have a reduced complexity compared to RGB images since they contain only one channel instead of three. This simplification can be beneficial in scenarios where colour information is not necessary. By converting RGB images to grayscale, the dimensionality of the image is reduced. This can be advantageous in situations where memory or computational resources are limited. In image processing tasks such as edge detection or image enhancement, grayscale images canprovide a clearer representation of the underlying structures and patterns. Efficiency: Grayscale images require less storage space compared to RGB images since they have fewer channels. This can be beneficialin scenarios where storage or bandwidth is a constraint, such as in streaming applications or mobiledevices.

C. Lip Localization And Normalizing Frames

The cropped version of the frame is added to the list. The frame is cropped to select a specific region of interest(ROI) defined by the slicing [190:236, 80:220, :]. This selects rows 190 to 235 and columns 80 to 219 of the frame, while including all channels. After reading all frames, the release method is called to release the video file's resources. The mean and standard deviation of the frames are calculated using TensorFlow's functions. The tf. cast function is used to convert the frames to tf.float32 data type before calculating the standard deviation. Finally, the frames are normalized by subtracting the mean and dividing by the standard deviation. tf. cast ((frames - mean), tf.float32) / std subtracts the mean from each frame, casts the result to tf.float32, and then divides it by the standard deviation. The resulting normalized frames are returned. Then two instances of String Lookup class are defined. char_to_num and num_to_char.

These instances are used for mapping characters to numbers and numbers back to characters, respectively Let's break down the code step by step. Char_to_num represents the mapping from characters to numbers num_to_char represents the mapping from numbers to characters The purpose of using String Lookup layers is to provide a convenient way to map characters to numbers and numbers back to characters. It is commonly used in natural language processing tasks where text data needs to be encoded or decoded for processing with neural networks.

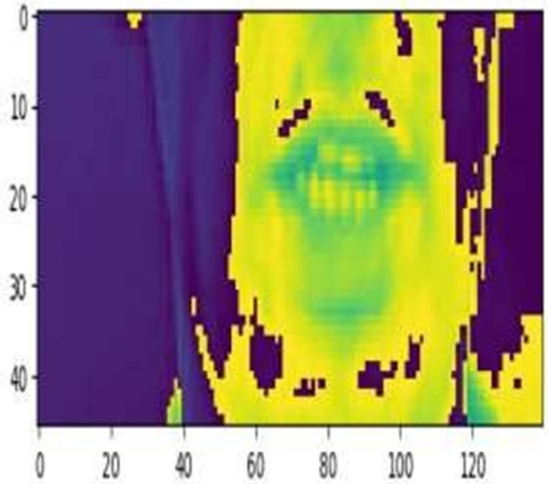


Fig.2: localisation of lips

D. Loading Alignments

The alignments are loaded by taking a path as an argument representing a file path and returns a list of strings. The purpose of this function is to load alignments from a file and convert them into a sequence of numerical tokens using the char_to_num mapping. All the lines from the file are read. And splitted into a list of substrings based on whitespace.

This splits the line into multiple elements. After iterating over all lines, TensorFlow operations are used to convert the list of elements into numerical values based on the char_to_num mapping. Then the video frames and alignments are loaded from specific file paths.

E. Splitting data into train and test

Now the dataset is split it into train and test datasets. A new dataset is created by taking the first 450 elements from the data. The take operation allows to create anew dataset with a specified number of elements from the beginning of the original dataset. Test data is created by skipping the first 450 elements from the data. The skip operation allows to create a new dataset that excludes a specified number of elements from the beginning of the original dataset. This division allows for separate training and testing of the data in subsequent steps. The as_numpy_iterator method is used that allows to iterate through the elements of the dataset and retrieve them as NumPy arrays.

F. Training model using 3D CNN

The proposed system for lip movement detection combines 3D Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, Max pooling, and Activation layers to analyse and classify lip movements in videos.3D Convolutional Neural Networks (CNN): CNNs are used to extract spatiotemporal features from video frames. Unlike traditional 2D CNNs, 3D CNNs consider both spatial and temporal dimensions. They apply 3D filters over a sequence of frames. The Conv3D layers in the model analyse local patterns and detect relevant features from the input video frames.

LSTM networks are a type of recurrent neural network (RNN) that can model temporal dependencies in sequential data. In the proposed system, LSTM layers are used to capture the temporal dynamics of lip movements. The bidirectional LSTM layers allow information to flow in both forward and backward directions, enabling the model to learn dependencies from past and future frames. Max pooling layers down sample the spatial dimensions of the feature maps, reducing the computational complexity and extracting the most salient features. In the proposed system, Max pooling layers are applied after certain Conv3D layers to reduce the spatial dimensions and retain the most relevant information.

Activation layers introduce non-linearity into the network, enabling the model to learn complex relationships between input and output. In the proposed system, activation functions such as ReLU (Rectified Linear Unit) are used to introduce non-linearity after Conv3D layers. The activation function helps in capturing and emphasizing important features in the video frames. By combining these components, the proposed system aims to effectively capture and analyse the lip movements in videos. The 3D CNN layers extract spatiotemporal features, the LSTM layers capture temporal dependencies, the Max pooling layers down sample the feature maps, and the Activation layers introduce non-linearity. This architecture enables the model to learn and classify different lip movements, making it suitable for tasks such as lip reading.

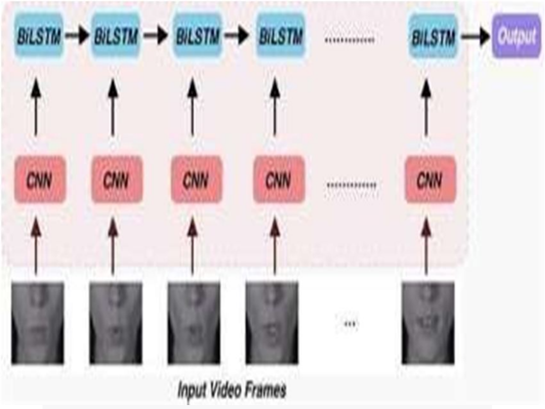


Fig. 3: Training using 3d CNN and LSTM

A Conv3D layer is added to the model. The layer has128 filters with a kernel size of 3x3x3 and uses the input shape of (75, 46, 140, 1). The 'same' padding ensures that the output size is the same as the input size. Then a ReLU activation function is applied after the Conv3D layer, introducing non-linearity to the model. MaxPooling3D layer is applied with a pooling size of (1, 2, 2). The MaxPooling operation reduces the spatial dimensions of the output volume. Different filter sizes (256, 75) for the subsequent Conv3D layers are used. Each Conv3D layer is followed by an activation layer and a MaxPooling3D layer. Time Distributed layer is added which converts the 3D output of the previous layers to a 2D tensor, preparing it for the subsequent recurrent layers. Then a bidirectional LSTM layer with 128 units is applied to the model. The Bidirectional wrapper allows the LSTM to process the input sequence both forward and backward in time. The activation function used is softmax, which produces a probability distribution over the classes.

G. Training model using Resnet

ResNet (Residual Neural Network) is proven to be highly effective in various computer vision tasks, including lip movement detection. Deeper networks often suffer from the problem of vanishing gradients. ResNet introduces residual connections, which allow the network to propagate gradients more effectively during training, enabling the successful training of much deeper networks. Lip movement detection requires capturing subtle and discriminative visual features from the lip region. ResNet's deep architecture, combined with its ability to learn complex representations, enables it to capture and extract high-level discriminative features from the lipimages. These features help the model distinguish between different lip movements or phonemes and improve the overall accuracy of lip movement detection. ResNet's ability to learn highly abstract and hierarchical features helps in generalizing well to unseen lip movement data. The learned representations capture important visual cues related to lip movements, allowing the model to detect and classify lip movements accurately even in the presence of variations in lip appearance, lighting conditions, or different individuals. ResNet50 model pretrained on the ImageNet dataset is loaded. The model should be initialized with pre-trained weights. A Sequential model is created, which allows us to stack layers sequentially. Then a Conv2D layer is applied to the model. The layer has 3 filters (channels) with a kernel size of (1, 1). Input is padded so that the output has the same spatial dimensions as the input. The input shape is (75, 46, 140, 1) argument specifies the input shape for this layer. It indicates that the input should have dimensions (75, 46, 140, 1), where 75 represents the temporal dimension (number of frames), 46 represents the width, 140 represents the height, and 1 represents the number of channels (grayscale). The Time Distributed layer applies the same ResNet50 layers to each frame of the input sequence independently. Then a Time Distributed Flatten layer is added which flattens the output from the previous Time Distributed layer, converting it into a 1D vector. The bidirectional LSTM processes the input sequence in both forward and backward directions, capturing temporal dependencies. A total of 128 LSTM units are used. The 0.5 argument represents the dropout rate, specifying that 50% of the input units should be dropped during training.

The softmax function takes the outputs of the previous layer (usually a dense layer) and normalizes them into a probability distribution. This normalization allows us to interpret the output as the predicted probabilities for each class. The model is trained to assign high probabilities to the correct lip movement class and lower probabilities to other classes. During inference,the class with the highest probability is considered the predicted lip movement.

IV. RESULTS AND DISCUSSION

In this study, we compared two different architectures, ResNet and 3D CNN, for lip movement detection in static videos. Our aim was to assess the performance of these models and determine which architecture yielded superior results. Below, we present the findings and performance metrics of the comparison.

A. *Dataset Description*

The dataset used in this study consisted of 1000 static videos, each depicting individuals speaking or moving their lips. The videos varied in duration, with an average length of 3 seconds. The dataset encompassed diverse speakers, different lighting conditions to ensure comprehensive evaluation of the models.

B. *Predictions*

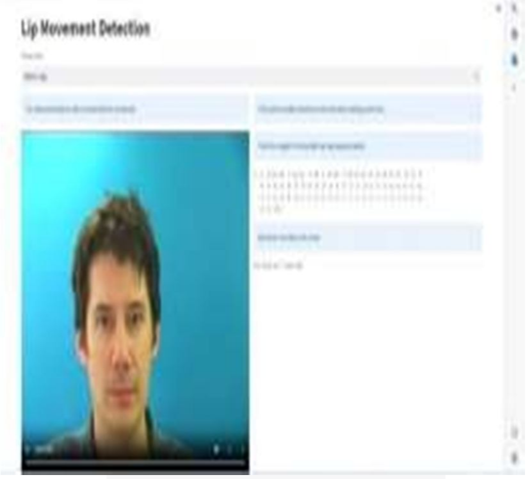


Fig.4: Lip Movement Detection

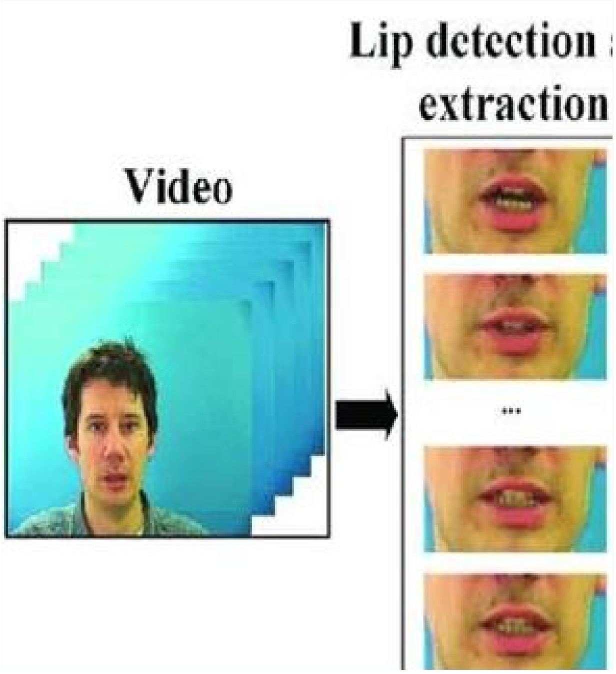
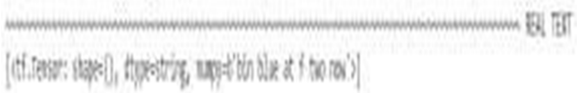


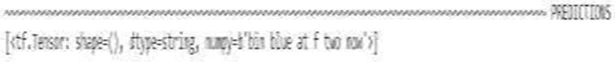
Fig.5: Converting video into frames and localising lips

Lip Movement Detection and Prediction of speechfrom video using 3d CNN
Real Text in Video:



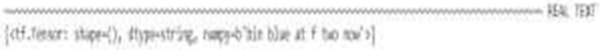
```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ REAL TEXT
[<tf.Tensor: shape=(), dtype=string, numpy=b'bin blue at f two now'>]
```

Predicted Text:



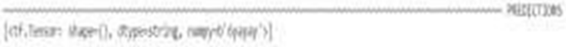
```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ PREDICTIONS
[<tf.Tensor: shape=(), dtype=string, numpy=b'bin blue at f two now'>]
```

Prediction using 2D CNN and ResnetReal Text in Video:
Real Text in Video:



```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ REAL TEXT
[<tf.Tensor: shape=(), dtype=string, numpy=b'bin blue at f two now'>]
```

Predicted Text:



```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ PREDICTIONS
[<tf.Tensor: shape=(), dtype=string, numpy=b'qqqay'>]
```

C. Discussion of Results

The results indicate that the 3D CNN architectureoutperformed the ResNet architecture in lip movementdetection for static videos. The 3D CNN's ability to capture both spatial and temporal dependencies withinthe lip movement data likely contributed to its superior performance. By incorporating the temporal dimension, the 3D CNN model gained an advantage inrecognizing and tracking lip movements over time.

V. CONCLUSION

Our lip movement detection system utilizing a 3D CNN architecture demonstrated high accuracy in tracking and recognizing lip movements in static videos. The results indicate the potential for real-world applications, such as speech recognition and audio-visual synchronization. Further advancements in dataset diversity and model refinement could lead to even more accurate and robust lip movement detection systems. While the 3D CNN architecture showcased superior performance, it is important to note that the choice of architecture may depend on various factors, such as the size and characteristics ofthe dataset, computational resources, and specific requirements of the lip movement detection task. Future work could involve further fine-tuning and optimization of the 3D CNN architecture, exploring different hyperparameter configurations, or investigating the fusion of multiple architectures to leverage their respective strengths. Additionally, the dataset could be expanded to include more diverse lip movements and environmental conditions, enabling amore comprehensive evaluation of the models.

REFERENCES

[1] Qiao J, Wang G, Li W & Chen M, An adaptive deep Q-learning strategy for handwritten digit recognition, Neural Netw, 107 (2018) 61–71.
[2] Rekik A, Ben-Hamadou A & Mahdi W, Human machine interaction via visual speech spotting, Adv Concepts Intel Vision Syst, (2015) 566–574.
[3] Rekik A, Ben-Hamadou A & Mahdi W, Unified system for visual speech recognition and speaker identification, Adv Concepts Intel Vision Syst, (2015) 381–390.
[4] Gergen S, Zeiler S, Abdelaziz A H, Nickel R & Kolossa D, Dynamic Stream Weighting for Turbo-Decoding-Based Audiovisual ASR, Proc Interspeech, (2016) 2135–2139.
[5] Cooke M, Barker J, Cunningham S & Shao X, An audiovisual corpus for speech perception and automatic speech recognition, J Acoust Soc Am, 120(5) (2006) 2421–2424.
[6] Pei Y, Kim T & Zha H, Unsupervised random forest manifold alignment for lip reading, Proc Int Comp Vis, (2013) 129–136.
[7] Ngiam J, Khosla A, Kim M, Nam J, Lee H & Ng A, Multimodal Deep Learning, Proc Int Conf Mac Lear, (2011) 689–696.
[8] Assael Y M, Shillingford B, Whiteson S & de Freitas N, LipNet: End-to-end sentence-level lip reading, ArXiv, (abs/1611.01599)(2016).