A Survey on an Efficient Visual Information Based Speech Recognition

Prof. Veena K R[1] Sricharan S[2] Prithvi Simha[3] Sharath Chandra E. S[4] Sunil R[5]
[1,2,3,4,5]Department of Computer Science and Engineering
[1,2,3,4,5]Sapthagiri College of Engineering, Bangalore, India Affiliated to Visveswaraya Technological University, Belagavi, India

Abstract— The survey based on the Efficient Visual Information Based Speech Recognition discloses several aspects about the hidden difficulties in human lipreading. It compares the Face Detection Algorithms, Techniques of Speech Reconstruction, Decoding and Language Model Integration and various Lip-Reading Techniques to provide an insight into building a more efficient lip-reading system and also an efficient method of speech reconstruction methods, by taking into consideration, all the drawbacks of the existing systems.

Keywords: CNN (Convoluted Neural Networks) LSTM (Long Short-Term Memory), OOV (Out-Of-Vocabulary), ROC (Receiver Operating Characteristic Curves), ROI (Region of Interest)

I. INTRODUCTION

Lipreading is the method of deducing speech by examining the movement of lips. In other words, it could be described as the process of decoding text from visual information generated by the speaker's mouth movement. The function of lipreading is based on information yielded by the context and knowledge of the language. Lipreading is also known as visual speech recognition and is a disputable task for humans, particularly in the absence of a frame of reference. Different words can produce professedly indistinguishable lip movements, hence the lipreading is a debatable predicament in the word level. Paradoxically, the professional lipreaders have not achieved high accuracy in word predications.

The above drawbacks led to the revolution of automated lipreading system, i.e. a machine that can analyze and read the movement of lips with a reasonable accuracy. Many advancements have been made in machine learning made automated lipreading systems. Some of the practical applications of automated lipreading are improved hearing aids, silent dictation in public places, security, speech recognition in noisy environments, biometric identification, silent movie processing etc. However, these models could not achieve the expected results. The traditional lipreading models were revolutionized by deep learning and deep neural networks with a large number of datasets for training.

Generally, Lip reading techniques comprises of four main stages: face detection, cropping module, feature extraction and text decoding. These methods have to be performed in sequence to achieve lipreading. Face detection involves distinguishing between faces and non faces, cropping module crops to the ROI, feature extraction helps in extracting the desired features

II. RELATED WORKS

A. Face Recognition using Viola Jones Algorithm

Face recognition is all about deducing the facial characteristics using a specialized algorithm. There has been two major divisions while applying the face detection methods: Using Viola jones algorithm and using the deep CNN techniques.

Prof. K. B. Pawar et al [1], speaks in favor of a novel method which is correspondence measure method and a combined permutation of Voila-Jones Face detection method and Eigen Face classification technique. An exhaustive search is performed using a sliding window of different sizes, aspect ratios, and locations in the Voila-Jones algorithm. The Eigen Face classification technique speaks in the favor of reconstruction of the face images using a standard Eigen Picture and a small collection of weights for each face. In this work, to detect the facial characteristics, the Viola-Jones algorithm is considered, where the facial features are extracted by the mean and the covariance matrix i.e. covariance matrix's The Eigen Vectors and Eigen Values are calculated and the feature vectors are obtained.

Sushil Kumar Paul et al [2], have given prime importance to a novel adaptive technique where the facial feature points such as eyes corners, nostrils, nose tip, and mouth corners in frontal view faces are automatically extracted i.e. by applying the cumulative distribution function approach by varying different threshold values. In the first step, for detecting the location of face and cropping the face region (with and without forehead region), the Viola-Jones algorithm is applied. By varying the different threshold values a new filtered face can be created in an adaptive way and by this method the cumulative distribution function of the cropped face region without forehead area. For the eyes and mouth filtered image and the contour algorithm for nose filtered image the simple linear search algorithm is implemented and the desired corner points can be extracted automatically.

B. Face Recognition using Deep CNN

Face detection using the Deep CNN is inconceivable because of its complexity.

Hamid Ouanan et al [5], speaks in the favor of assembling a novel dataset with a colossal number of faces labelled for identity by deploying a smart synthesis augmented approach based on rendering pipeline to augment the pose and lighting variability. In addition to this a robust deep CNN model is used with which a new real time application of this approach proposed. The proposed application is called PubFace, which allows users to identify anyone in public spaces. The VGGNet, off-the-shelf deep models of, originally trained on the ImageNet, large-scale image recognition benchmark (ILSVRC) is used for this purpose. The CNN model is fine-tuned on the dataset. The LFW dataset is experimented to evaluate the implementation of the advocated approach i.e. a juxtaposition is made with the proposed approach with the competitive supervised methods and current best commercial system. The Receiver Operating Characteristic Curves (ROC) are used to judge the performance of the advocated approach.

Alexandros Koumparoulis et al [6], speaks about exploring the ROI size. Region of interest (ROI) in other words is simply the speaker's mouth. Visual feature extraction is an important part of lip-reading systems, while also shared by a number of associated audio-visual processing tasks. Since the research focusses on face detection they have worked on CNN based face detection and landmark localization in which 16 main facial landmarks are marked for Landmark post-processing and ROI extraction where the CNN predictions (obtained independently at each frame) are subjected to median filtering over a 15-frame window. Their work mainly was achieved with the use of LSTM (Long short-term memory)

S. NO	Face detection algorithms	Precision	Recall
1.	Viola Jones based face detector[1][2][11]	0.27321	0.27321
2.	Deep CNN based Face detector[5][6][11]	0.339450	0.037582

We notice from the comparison that recall or the true positive rate of Viola Jones Face detector algorithm is higher and hence it has been widely popularized and used.

C. Reconstruction of Speech

Speechreading is the task of inferring phonetic information from visually observed facial movements, and is a notoriously difficult task for humans to perform. It is closely related to Lipreading.

Ariel Ephrat et al [4], have given top priority for reconstruction of speech from a silent video by significantly enhancing state-of-the-art reconstructed speech's digestibility by using an end-to-end CNN-based model that is capable of predicting the speech audio signal of a silent video of a person speaking. It is also evident that if the model is allowed to learn from the speaker's entire face rather than the mouth region alone, significantly improves the performance. It is significant that the Out-Of-Vocabulary (OOV) words can be reconstructed by allowing the speechreading method to be modelled as a regression problem. Sound features are generated for each frame i.e. based on the neighboring frames by the proposed CNN model. Now, from the learned speech features, by synthesizing the waveforms intelligible speech can be produced. Gaussian white noise is used as the excitation signal which helps in improving the feasibility of reconstructing an intelligible audio speech signal from silent videos frames.

D. Decoding and Language Model Integration

Decoding involves processing the data and obtaining the desired results.

Jan Chorowski et al [10], have proposed practical solutions to problems like overconfidence in predictions and a tendency to produce incomplete transcripts when language models are used. The listener which converts frames to hidden activations, the speller and attention mechanism, training criterion, beam search and language model integration works hand in hand to achieve the desired results. However, one drawback was with respect to generation of partial transcripts. The main advantage of this system was that seq2seq networks were locally normalized, i.e. every step probability distribution is produced by the speller. Alternatively, normalization can be performed globally on whole transcripts.

E. Lip Reading Techniques Worked upon till Date

Lipreading is the task of decoding text from the movement of a speaker's mouth, it is more complex and technically inviable than it sounds. It requires the model to be trained from end to end in order to achieve lipreading.

Michael Wand et al [3], have worked on a lipreading system which yields an end-to-end trainable system which consumes an infinitesimal number of frames of untranscribed target data to revamp the recognition accuracy on the target speaker by using domain-adversarial training for speaker independence which is integrated into the lipreader's advancement based on a stack of feedforward and LSTM (Long Short-Term Memory) recurrent neural networks. The main goal is to push the network to learn an intermediate data representation which is domain-agnostic i.e. it should be independent whether input data is obtained from target speaker or a source speaker. TensorFlow's Momentum Optimizer is applied using the stochastic gradient descent in order to minimize the multi-class cross-entropy hereby achieving optimization.

Brendan Shillingford et al [7], have given prime importance to transforming a raw video into a word sequence. The first component of this system is a data processing pipeline used to create the Large-Scale Visual Speech Recognition (LSVSR) dataset used in this work, distilled from YouTube videos and comprising of phoneme sequences coupled with video clips of faces speaking. Their approach was first to combine a deep learning-based phoneme recognition model with production-grade word-level decoding techniques. By decoupling phoneme prediction and word decoding as is often done in speech recognition, hence it is possible to arbitrarily extend the vocabulary without retraining the neural network.

Lele Chen et al [8], have worked mainly on the lip movements detection. They have taken speech audio and a lip image of the target identity as input, and generates multiple lip images (16 frames) in a video depicting the corresponding lip movements. Observing that speech is highly correlated with lip movements even across identities, a concept grounds lip reading, the core of their paper is Lip Movements Generation at a Glance. To explore the best modelling of such correlations in building and training a lip movement generator network. They devised a method to fuse time-series audio embedding and identity image embedding in generating multiple lip images, and propose a novel audio-visual correlation loss to synchronize lip changes and speech changes in a video.

Joon Son Chung et al [9], have used the recent sequence-to-sequence (encoder-decoder with attention) translator architectures that have been developed for speech recognition and machine translation. In this paper the dataset developed is established from thousands of hours of BBC television broadcasts which have speaking faces along with subtitles of what is being said. Their model is devised in such a way that it can operate over dual attention mechanism that can operate over visual input only, audio input only, or both. They have an image encoder, audio encoder and character decoder in place to achieve what is called lipreading. With or

without the audio the goal was to recognize the phrases spoken by the talking face[9].

III. CONCLUSIONS

In terms of face recognition and detection techniques, it is evident that Viola-Jones Algorithm is comparatively more efficient and hence feasible for implementation [11]. Speech Reconstruction Techniques helps to consider the OOV (Out-of-Vocabulary) words for better optimization. Decoding Techniques helps to overcome overconfidence in predictions and tendency to produce incomplete transcripts. A survey on Lipreading Techniques employed till date helps to understand and develop a system which tries to overcome the existing drawbacks like overconfidence in prediction, high false positive rate and improper decoding due to noise[12][13][14].

REFERENCES

[1] Prof. K. B. Pawar, Prof. Feroza Mirajkar, Prof. Vinod Biradar, Dr Ruksar Fatima, "A Novel Practice for Face Classification". International Conference on Current Trends in Computer, Electrical, Electronics and Communication (ICCTCEEC-2017).
[2] Sushil Kumar Paul, Mohammad Shorif Uddin, Saida Bouakaz, "Extraction of Facial Feature Points Using Cumulative Distribution Function by Varying Single Threshold Group". IEEE/OSA/IAPR International Conference on Infonnatics, Electronics & Vision 15 Mar 2012.
[3] Michael Wand and Jurgen Schmidhuber, "Improving Speaker-Independent Lipreading with Domain-Adversarial Training". The Swiss AI Lab IDSIA, USI & SUPSI, Manno-Lugano, Switzerland, arXiv:1708.01565v1 [cs.CV] 4 Aug 2017.
[4] Ariel Ephrat and Shmuel Peleg, "VID2SPEECH: Speech Reconstruction from Silent Video". IJCSI International Journal of Computer Science Issues, arXiv:1701.00495 [cs.CV] 9 Jan 2017.
[5] Hamid Ouanan, Mohammed Ouanan, and Brahim Aksasse, "Face Recognition Using Deep Features". Springer International Publishing AG 2018 M. Ezziyyani et al. (eds.), Advanced Information Technology, Services and Systems, Lecture Notes in Networks and Systems 25, https://doi.org/10.1007/978-3-319-69137-4_8 12 Nov 2017.
[6] Alexandros Koumparoulis, Gerasimos Potamianos, Youssef Mroueh and Steven J. Rennie, "Exploring ROI size in deep learning based lipreading". Electrical and Computer Eng. Dept., University of Thessaly, Volos 38221, Greece Athena Research and Innovation Center, Maroussi 15125, Athens, Greece IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, U.S.A.
[7] Brendan Shillingford, Yannis Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Ben Coppin, Ben Laurie, Andrew Senior and Nando de Freitas, "LARGE-SCALE VISUAL SPEECH RECOGNITION". DeepMind & Google.
[8] Lele Chen, Zhiheng Li, Ross K. Maddox, Zhiyao Duan and Chenliang Xu, "Lip Movements Generation at a Glance". Wuhan university and University of Rochester
[9] Joon Son Chung, Andrew Senior, Oriol Vinyals and Andrew Zisserman, "Lip Reading Sentences in the Wild". Department of Engineering Science, University of Oxford 2Google DeepMind.
[10] Jan Chorowski and Navdeep Jaitly, "Towards better decoding and language model integration in sequence to sequence models". Google Brain Google Inc. Mountain View, CA 94043, USA.
[11] Kriti Dang, Shanu Sharma, CSE Dept. ASET, AMITY University, "Review and Comparison of Face Detection Algorithms".
[12] Christoph Bregler and Stephen M Omohundro, "Learning Visual Models for Lipreading".
[13] Kai Xu, Dawei Li, Nick Cassimatis and Xialong Wang, Arizona State University, "LCANET: End-to-End Lipreading with Cascaded Attention- CTC".
[14] Leon Rothkrantz. "Lip reading using Surveillance Cameras".
[15] Gregory j Wolff, K Venkatesh Prasad, David G Stork and Marcus Hennecke. "Lipreading by Neural Networks: Visual Pre-processing, learning and Sensory Integration"

