DOI: 10.47813/10.47813/ MIP: Engineering-IV-2022.4. 82-90

EDN: EONWEX

End-to-end Visual Speech Recognition for Human-Robot Interaction

Denis Ivanko*, Dmitry Ryumin, Maxim Markitantov

St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS)

E- mail: Ivanko.d@iias.spb.su

Abstract. In this paper we present a novel method designed for word-level visual speech recognition and intended for use in human-robot interaction. The ability of robots to understand natural human speech will significantly improve the quality of human-machine interaction. Despite outstanding breakthroughs achieved in this field in recent years this challenge remains unresolved. In current research we mainly focus on the visual part of the human speech, so-called automated lip-reading task, which becomes crucial for human-robot interaction in acoustically noisy environment. The developed method is based on the use of state-of-the-art artificial intelligence technologies and allowed to achieve an incredible 85.03% speech recognition accuracy using only video data. It is worth noting that the model training and testing of the method was carried out on a benchmarking LRW database recorded in-the-wild, and the presented results surpass many existing achieved by the researchers of the world speech recognition community.

Keywords: visual speech recognition, robot, automated lip-reading task

1. Introduction

Automatic speech recognition is a field of growing attention in modern human-robot interaction. It is a natural form of communication and the ability of robots to understand human speech is difficult to overestimate. Such ability has already proved its usefulness in many practical applications [1]. However, in acoustically noisy conditions, such as public places, construction environment, road traffic, etc. audio speech recognition for robots doesn't perform at its best. From this perspective, visual speech recognition (or so-called automated lip-reading) is a natural complement to audio-based speech recognition [2]. It can facilitate speech recognition in aforementioned noisy environments.

Traditional speech recognition system consists of two stages: features extraction block and speech recognition block [3]. Very recently, an end-to-end approach for visual speech recognition have been presented [4]. Its core idea is to use only one neural network which perform the functions of both blocks: (1) extract informative features and (2) speech recognition. However, research on end-to-end lip-reading has been very limited. The proposed in this paper method and developed visual speech recognition system belongs to end-to-end category.

Motivated by these problems, we develop a method for word-level visual speech recognition and build an end-to-end model to be used in modern human-robot interaction systems. The developed technologies and the obtained results are in line with the state-of-the-art world level and surpass it in some aspects.

In summary, we make three main contributions. Firstly, we propose a novel method for word-level visual speech recognition to be used in human-robot interfaces. Secondly, we build from scratch state-of-the-art end-to-end model and train it on a real-world data. Finally, we report the evaluation results on benchmarking LRW dataset and compare it with existing methods.

The rest of the paper is organized as follows. In Section 2, we refer to recent works on automated lip-reading, with emphasis on those that apply end-to-end methods. In Section 3, we discuss benchmarking LRW dataset recorded in-the-wild conditions. In Section 4, we present the proposed method and describe the architecture of the developed end-to-end model. Finally, In Section 5, we present our evaluation results along with the comparison to the existing methods.

2. Related works

Prior to the appearance of modern end-to-end and deep learning approaches, most of the work in visual speech recognition for human-robot interfaces was based on hand-crafted features and modelled by Hidden Markov model pipeline [5]. However, in recent years researchers starting to use deep learning methods either for extracting informative features [6-8] or for building end-to-end architectures [9-12].

Perspectives of using automated lip-reading technologies for robotic control or in human robot interfaces were considered in our previous work [13]. In general, visual speech recognition for robots has a long history. According to the design of the front end network, the current state-of-the-art visual speech recognition methods can be divided into three main categories: 2-dimensional (2D) convolutional neural networks (CNN), such in [14], 3-dimensional convolutional neural networks (3D CNNs), such in [15], or a combination of 2D and 3D convolutions, which inherit the advantages of both [16]. Recently, method of third type have become widely used in visual speech recognition due to its ability to simultaneously capture temporal dynamics of lips movements and extract discriminative features.

For sequence modeling long-short term memory (LSTMs) or its variations are often used [17]. When temporal modeling is required, such in visual speech recognition, LSTMs usually lead to better performance commonly used in NLP, video-prediction, automated lip-reading, etc [18-20].

In this paper, we develop own method which is based on the combination of different state-of-the-art artificial intelligence technologies and improve upon earlier results by achieving around 85% speech recognition accuracy with 500 recognition classes using only video data.

3. Dataset

This section describes a used in research large-scale visual speech recognition dataset LRW [21]. The dataset consists of the recordings of 500 English words, forming up to 1000 difference utterances and pronounced by hundreds of speakers. All videos are 29 frames in length. The LRW main characteristics are presented in the table 1.

Table 1. LRW dataset characteristics.

Dataset	Classes	Samples for each class	Number of frames
Train	500 (words)	800-1000	29
Valid		50	
Test		50	

These 500 words occur more than 800 times in the training set and more than 40 times in validation and test sets. Worth noting, that the words are not isolated: they are taken in-the-wild conditions, so some co-articulation of the lips from preceding and subsequent words is present. Snapshots of speakers of the LRW dataset are given in the figure 1.

Figure 1. Snapshots of speakers in LRW dataset.

4. Methodology

The functional diagram of the proposed method for visual speech recognition is shown in Figure 2. It consists of several sequential steps. In the preprocessing phase, we discard redundant information in order to focus on the lips region.

Firstly, at each video frame the mouth region is detected and cropped by using FaceMesh MediaPipe algorithm. The mouth region exaction procedure described in great detail in our previous work [22]. Secondly, before entering our end-to-end model, some augmentation procedures are performed, namely: (1) Grey-scale transformation, followed by (2) image normalization and (3) histogram alignment. Then, to reduce overfitting and introduce less confidence in the predictions, the (4) MixUp data augmentation technique was applied to images with a probability of 40% (only during training procedure). The merging ratio of the two images varied from 30 to 70% so that the sum was always 100% (zero transparency). Label Smoothing was applied to the labels of those images that did not have MixUp. Thirdly, the resulting images are formed into batches with Sequence_Length of 29 frames (LRW dataset dimension) and fed into end-to-end network for speech recognition.

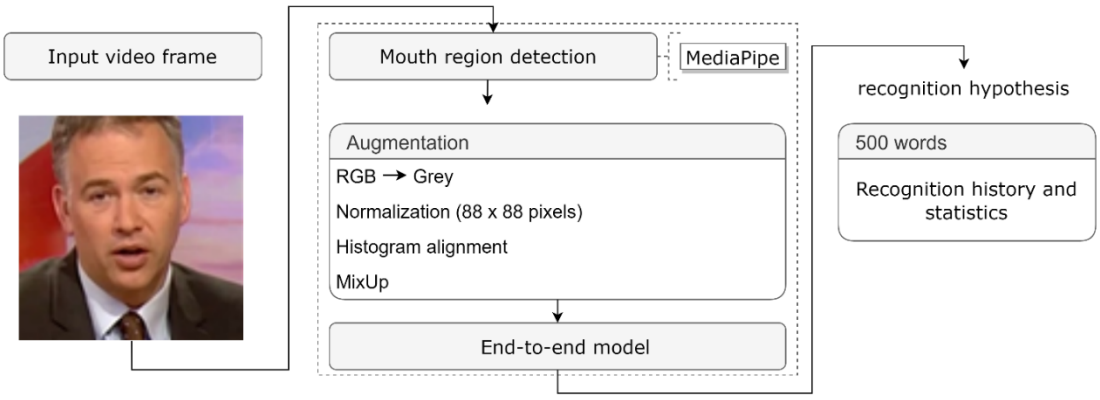

Figure 2. Functional diagram of the proposed visual speech recognition method.

The general architecture and layers dimension of the developed end-to-end model is shown in Table 2. To extract informative visual features, a modified 3DResNet-18 neural network was used with the addition of the Squeeze-and-Attention module (left column). ResNet makes it relatively easy to increase accuracy by increasing depth, which is more difficult to achieve with other networks. We use Squeeze-and-Attention modules as heads to extract features to fully exploit their multi-scale.

The back-end of the model is a Bidirectional LSTM network. The extracted features were applied to 2 layers of BiLSTM, 512 neurons each. The output of the first BiLSTM layer is sequence-to-sequence. The output of the second BiLSTM layer is sequence-to-one (right column). The last fully connected layer determines the most probable hypothesis from 500 recognition classes. The training process used a learning rate scheduler technique - cosine annealing.

We should emphasize that the overall visual speech recognition system can be directly trained end-to-end. Proposed method for word-level visual speech recognition is based on recent advances in deep learning and apply them to automated lip-reading to be used in human-robot interfaces.

Table 2. End-to-end visual speech recognition model architecture

1) Batch of video frames (Sequence_Length × 1 ×88 × 88)	7) Global average pooling (Sequence_Length × 512)
2) 3D Conv (Sequence_Length × 64 × 44 × 44)	8) Dropout layer, p = 0.4

3) SA-Residual Block × 2 (Sequence_Length × 64 × 44 × 44)	9) BiLSTM + Dropout, p = 0.2 (Sequence_Length × 2 × 1024)
4) SA-Residual Block × 2 (Sequence_Length × 128 × 22 × 22)	10) BiLSTM + Dropout, p = 0.2 (1024)
5) SA-Residual Block × 2 (Sequence_Length × 256 × 11 × 11)	11) FC + Softmax (500)
6) SA-Residual Block × 2 (Sequence_Length × 512 × 6 × 6)	

5. Evaluations

In this section we present the recognition results of developed visual speech recognition method and end-to-end neural network architecture on benchmark LRW dataset. The comparison of recognition accuracy with and without MixUp augmentation technic are shown in Table 3. The application of MixUp results in increase of recognition accuracy to 85.03% (2,83% absolute).

Table 3. Comparison of recognition results of our model with / without MixUp augmentation.

№	Neural network model architecture	Recognition accuracy
1	MPipediae FaceMesh + Label Smoothing + Squeeze-and-Attention + 3DResNet-18 + BiLSTM + Cosine WR	82.2%
2	MediaPipe FaceMesh + Label Smoothing + MixUp + Squeeze-and-Attention + 3DResNet-18 + BiLSTM + Cosine WR	85.03%

The comparison of recognition results of our model with some state-of-the-art approaches are shown in Table 4. As we can see from the table, our model outperforms recent state-of-the-art approaches up to 1.73% absolute if compare with work [12]. Smallest gap in recognition accuracy seen in comparison with recent research work [5], only 0.23%. However, with 500 recognition classes and sufficient amount of test data in LRW dataset, this result of our model is a significant improvement upon previous approaches, confirming the viability of the proposed method and model.

Table 4. Comparison of recognition results of our model with state-of-the-art on LRW dataset.

№	Neural network model architecture	Recognition accuracy
1	3D Conv + ResNet-34 + Bi-LSTM [12]	83.30%
2	Multi-grained + Bi-ConvLSTM [11]	83.34%
3	3D Conv + ResNet-34 + Bi-GRU [10]	83.39%
4	PCPG [9]	83.50%

5	DFTN [8]	84.13%
6	SpotFast + Transformer + Product-Key memory [7]	84.40%
7	3D Conv + ResNet-18 + Bi-GRU [6]	84.41%
8	3D Conv + P3D-ResNet50 + TCN [5]	84.80%
9	Our model	85.03%

6. Conclusion

In this paper we introduce an end-to-end visual speech recognition method designed for word-level automated lip-reading and intended for use in human-robot interaction. The ability of robots to understand natural human speech will significantly improve the quality of human-machine interaction. The developed method is based on the use of state-of-the-art artificial intelligence technologies. Proposed methodology allows to achieve 85.03% speech recognition accuracy using only video data. It is worth noting that the model training and testing was carried out on a benchmarking LRW database recorded in-the-wild conditions, which is currently one of the largest visual speech datasets in the world. Some possible future directions include audio-visual modalities fusion and implementation of developed technologies in interface of mobile information robot.

Acknowledgments

This research is financially supported by the Russian Science Foundation (project No. 21-71-00132).

References

1. Dalu, F. "Learn an Effective Lip Reading Model without Pains" / F. Dalu, S. Yang, S. Shan and X. Chen // In arXiv preprint arXiv:2011.07557. – 2020. – P. 1-6.
2. Kim, M. "Multi-modality associative bridging through memory: Speech sound recollected from face video" / M. Kim, J. Hong, S. J. Park, Y. M. Ro // In Proceedings of the IEEE/CVF International Conference on Computer Vision. – 2021. – P. 296-306.
3. Martinez, B. "Lipreading using temporal convolutional networks" / B. Martinez, P. Ma, S. Petridis, M. Pantic // In ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2020. – P. 6319-6323.
4. Zhang, Y. "Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition" / Y. Zhang, S. Yang, J. Xiao, S. Shan and X. Chen // In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition. – 2020. – P. 356-363.

5. Xu, B. "Discriminative multi-modality speech recognition" / B. Xu, C. Lu, Y. Guo and J. Wang // In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2020. – P. 14433-14442.
6. Zhao, X. "Mutual information maximization for effective lip reading" / X. Zhao, S. Yang, S. Shan and X. Chen // In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition. – 2020. – P. 420-427.
7. Wiriyathammabhum, P. "SpotFast Networks with Memory Augmented Lateral Transformers for Lipreading" / P. Wiriyathammabhum // In International Conference on Neural Information Processing. – 2020. – P. 554-561.
8. Xiao, J. "Deformation flow based two-stream network for lip reading" / J. Xiao, S. Yang, Y. Zhang, S. Shan and X. Chen // In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition. – 2020. – P. 364-370.
9. Luo, M. "Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading" / M. Luo, S. Yang, S. Shan and X. Chen // In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition. – 2020. – P. 273-280.
10. Petridis, S. "End-to-end audiovisual speech recognition" / S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, M. Pantic // In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). – 2018. – P. 6548-6552.
11. Wang, C. "Multi-grained spatio-temporal modeling for lip-reading" / C. Wang // In arXiv preprint arXiv:1908.1161. – 2019.
12. Stafylakis, T. Combining residual networks with LSTMs for lipreading" / T. Stafylakis, G. "Tzimiropoulos // In arXiv preprint arXiv:1703.04105. – 2017.
13. Ivanko, D. "An Experimental Analysis of Different Approaches to Audio–Visual Speech Recognition and Lip-Reading" / D. Ivanko, D. Ryumin, A. Karpov // In Proceedings of 15th International Conference on Electromechanics and Robotics" Zavalishin's Readings. –2021. – P. 197-209.
14. Ivanko, D. "Developing of a Software–Hardware Complex for Automatic Audio–Visual Speech Recognition in Human–Robot Interfaces" / D. Ivanko, D. Ryumin, A. Karpov // In Electromechanics and Robotics. – 2022. – P. 259-270.

15. Verkhodanova, V. "HAVRUS corpus: high-speed recordings of audio-visual Russian speech" / V. Verkhodanova, A. Ronzhin, I. Kipyatkova, D. Ivanko, A. Karpov, M. Železný // In International Conference on Speech and Computer. – 2016. – P. 338-345.
16. Ryumina, E. "A Novel Method for Protective Face Mask Detection using Convolutional Neural Networks and Image Histograms" / E. Ryumina, D. Ryumin, D. Ivanko, A. Karpov // In International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. – 2021. – P. 177-182.
17. Kashevnik, A. "Multimodal Corpus Design for Audio-Visual Speech Recognition in Vehicle Cabin" / A. Kashevnik, I. Lashkov, A. Axyonov, D. Ivanko, D. Ryumin, A. Kolchin, A. Karpov // In IEEE Access. – 2021. – № 9. – P. 34986-35003.
18. Shillingford, B. "Large-scale visual speech recognition" / B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, N. de Freitas // In arXiv preprint arXiv:1807.05162. – 2018.
19. Afouras, T. "LRS3-TED: a large-scale dataset for visual speech recognition" / T. Afouras, J. S. Chung, A. Zisserman // In arXiv preprint arXiv:1809.00496.
20. Zhu, H. "Deep audio-visual learning: A survey" / H. Zhu, M. D. Luo, R. Wang, A. H. Zheng, R. He // In International Journal of Automation and Computing. – 2021. – P. 1-26.
21. Chung, J. “Lip reading in the wild” / J. Chung, A. Zisserman // In Asian conference on computer vision. – 2016. – P. 87-103.
22. Ivanko, D. "Development of Visual and Audio Speech Recognition Systems Using Deep Neural Networks" / D. Ivanko, D. Ryumin // In International Conference Graficon. – 2021. – P. 1-12.