Lip Reading using CNN Lip Deflection Classifier and GAN Two-Stage Lip Corrector

Chengwei ZHANG, Senyao ZHANG
School of Artificial Intelligence, Hebei University of Technology, Tianjin 300132, China
184948@stu.hebut.edu.cn

Abstract. Lip reading is a task to infer the content of speech through the movement of the lips, which is a technology with wide application prospects. Thanks to the development of deep learning technology and the proposal of large-scale datasets, lip reading has made great progress. However, these large-scale datasets contain only small angle lip deflection which is different with the wild environment. Most networks based on these datasets take some simple methods to correct the lip deflection that led to performance degradation when in the wild. In this work, we proposed a very challenging dataset with large angle lip deflection and multiple speakers to simulate the wild environment, and we designed a lip deflection classifier based on Convolutional Neural Network and a two-stage corrector based on Generative Adversarial Network. We use these two models to correct the lip deflection and improve the recognition accuracy. Our proposed method achieves an absolute improvement of 18.3% and 7.4%, respectively, compared with no preprocessing and just face alignment, which shows that our proposed network is more suitable to correct the large angle lip deflection in the wild.

1. Introduction
Lip reading is a complex task that combines computer vision and natural language processing. It can be used to automatically infer the text content of visual information, and has a wide range of applications, such as speech recovery from silent surveillance videos or movies.

In recent years, the development of lip reading benefiting much from the following two aspects. The first aspect is the rapid development of deep learning technology, derived from neuroscience, which has achieved great success in areas such as image processing [1] and language modelling [2]. The second aspect is the proposal of large-scale datasets [3], which provides a lot of training data and complex environmental changes for lip reading and has greatly promoted the development of lip reading.

Most researchers use end-to-end models, including front-end feature extraction model and back-end classification model. The front-end feature extraction model is mainly responsible for extracting the features of single frame. The back-end classification model is mainly responsible for learning the pattern changes of the whole sequence from the feature vectors generated by the front-end model.

The accuracy of front view lip reading is always higher than the profile lip reading, which mainly because the front view contains more information than the profile. Some researchers use face alignment [4], combined with conventional data augmentation methods such as horizontal flipping, random cropping, and contrast enhancement [1]. These results all show that the aligned face and lip correction is helpful to improve the accuracy of lip reading. But these methods are all aimed at solving the deflection of the face on the roll and ignored the deflection on yaw.
In this paper, we solve the problem that the recognition accuracy is influenced by the large angle lip deflection. We proposed a two-stage lip corrector, which is a lip region correction model based on Generative Adversarial Network [5]. Firstly, we use the lip deflection classifier to estimate the lip deflection angle, and then feed it into the two-stage lip corrector to get the $0^\circ$ lip image. The lip corrector is divided into two stages. In the first stage, the lip image in the small angle range near the conversion point is firstly converted to the conversion point. In the second stage, the image at the conversion point generated by the first stage is converted to the $0^\circ$ lip image. After the two-stage lip corrector, we correct the lip deflection on yaw, which is effective for extend the information that the lip image contains and increase the recognition accuracy.

2. Architecture

2.1. CNN lip deflection classifier

2.1.1. AlexNet

AlexNet [1] is proposed by Hinton, which is the first deep learning network combine CNN with ReLU, Dropout, LRN and so on. AlexNet has 60 million parameters and 65,000 neurons, five layers of convolution, three layers of fully connected network, and the final output layer is the 1000-channel SoftMax.

2.1.2. GoogleNet

GoogleNet [6] is a powerful network in image classification. The network is based on inception structure, which uses the dense component to approximate the optimal local sparse structure.

2.1.3. ResNet

ResNet [7] is a very deep network which uses residual connection to make the training process easier. The network contains five different depth version named ResNet18, ResNet34, ResNet50, ResNet101, ResNet152.

2.2. GAN two-stage lip corrector

2.2.1. First-stage

The first-stage consists of seven first-stage sub-correctors, numbered 1 through 7. Each first-stage sub-correctors contains a conversion point with an angle of $\theta_i^*$, which is responsible for converting the lip image in range $\theta_i^* \pm 10^\circ$ to $\theta_i^*$ where $i$ represent the number of correctors, $\theta$ represent the angle of the corrector conversion point. The relationship between the angle range of the first-stage sub-correctors and the number is as follows.

$$\begin{align*}
-90 + 20 \cdot i \leq \theta_i &\leq -70 + 20 \cdot i \\
-10 &\leq \theta_i \leq 10 \\
-90 + 20 \cdot i \leq \theta_i &\leq -70 + 20 \cdot i
\end{align*}$$

(1)

where

$$\begin{align*}
i &\leq 3 \\
i &\leq 4 \\
i &\leq 7
\end{align*}$$
2.2.2. Second-stage

The second-stage contains six second-stage sub-correctors. Since the lip image can be directly obtained from the 0° lip image through the No. 4 first-stage sub-corrector, there is no need for the No. 4 second-stage sub-corrector, so it is numbered 1-3,5-7. For each second-stage sub-corrector, the image corrected to the conversion point obtained by the corresponding numbered first-stage sub-converter will be fed into the corresponding second-stage sub-correctors and the output is converted to the image of 0°.

2.3. Lip reading network

We adopt a traditional end-to-end lip recognition network architecture, which consists of the front-end feature extractor and back-end recognition network. Our proposed network uses 3DConv+ResNet18 as front-end. Firstly, the first 3D convolution layer is used to initially extract the features, and then the sequence are fed into the ResNet18 network for further feature extraction and finally generate a tensor of size B×L×1024, where B is Batch size, L is sequence length. For the back-end, we use the MS-TCN [8] which is a variation of TCN [9]. In TCN, the kernel size is same in each layer, but MS-TCN has different kernel size in each layer. To realize the multi-scale convolution, they improved the structure of convolution block, proposed a multi-branch network which has different kernel size in every branch, which makes the size of perception field varying in each layer.
3. Training data

3.1. Lip reading dataset
We collect a lip reading dataset of 10 subjects, 5 males and 5 females, with a total of 10 phrases, which were partially consistent with the phrases in Ouluvs2 [10]. Each subject read the same phrase three times and used a mobile HD camera to record the video with deflections.

3.2. 3D face dataset
For controlling the angle of face image, we use the FaceGen 3D face modelling software to model several different subjects, we set the lip deflection angle range of $\pm 70^\circ$ and use the dlib toolkit [11] to locate the lip region. Interval of $1^\circ$ exporting images, we get $141^\circ$ lip image, each category contains 5000 images, to get the raw data. We divide these images into two datasets, namely the angle classification dataset and the two-stage lip corrector dataset, which are organized in different ways.

3.3. Angle classification dataset
The angle classification dataset randomly selects 10% data from the 3D face dataset, stores each angle as a classification. This dataset is used to train the CNN angle classifier.

3.4. Two-stage lip corrector dataset
The lip corrector dataset is divided into a first-stage corrector dataset and a second-stage corrector dataset. The data of the first-stage corrector contains 7 subsets, each of them contains the lip image of two domains A and B. The A domain contains the lip image of $\theta_1 \pm 10^\circ$, and the B domain contains the lip image of $\theta_2^\circ$. The first-stage corrector is trained by using the same numbered subsets of the dataset of the first-stage corrector. The second-stage corrector contains 6 subsets, each of which contains the lip image of two domains A and B. The A domain contains the lip image of degree $\theta_3^\circ$, and the B domain contains the lip image of $0^\circ$. The second-stage corrector is trained by using the same numbered subsets of the second-stage corrector dataset.

4. Experiment and results

4.1. Classifier experiment
The training of classifier and two-stage corrector can be carried out separately. We first introduce the experimental results of classifier. We use Adam to optimise the network and the initial learning rate is $1e^{-4}$ which will reduce to by half for every 10 epochs. We train it for 30 epochs. We give the classification result matrices of Resnet18, GoogleNet and Alexnet deflection classifiers, and compare the classification accuracy.

![Fig.3 Classification result of 3 different CNN classifier](image)

|       | AlexNet | GoogleNet | ResNet18 |
|-------|---------|-----------|----------|
| **Top-1** | 0.60    | *0.67*    | 0.42     |
| **Top-3** | 31.41   | 36.77     | *48.93*  |
| **Top-5** | 50.14   | 60.67     | *76.48*  |
4.2. Two-stage lip corrector experiment
Two-stage lip corrector are based on GAN structure. We use different GAN such as Pix2PixGAN [12], CycleGAN [13] and StarGAN [14] as first-stage corrector or second-stage corrector. We train each GANs with different settings which is same as it first proposed. We also compared the different combination of them. The generation quality is evaluated by SSIM index.

Table 2. Result of each stage.

|               | Pix2PixGAN | CycleGAN | StarGAN |
|---------------|------------|----------|---------|
| **First-stage** | 0.67       | 0.71     | 0.79    |
| **Second-stage** | **0.87**   | 0.82     | 0.84    |

Table 3. Result of different combination.

| First-stage   | Second-stage | SSIM |
|---------------|--------------|------|
| Pix2PixGAN    | Pix2PixGAN   | 0.62 |
| Pix2PixGAN    | CycleGAN     | 0.63 |
| CycleGAN      | StarGAN      | 0.69 |
| **StarGAN**   | Pix2PixGAN   | **0.76** |

4.3. Lip reading experiment
To prove the superiority of our two-stage lip corrector, we compare with the traditional face alignment method, and the recognition network adopts the network structure of ResNet18+MS-TCN. We use Adam to optimise the network and the initial learning rate is 1e-6 and use the cosine learning rate scheduler to adjust the learning rate. We train it for 200 epochs. Experimental results show that our two-stage lip corrector plays a significant role in improving the recognition accuracy. Our proposed method achieves an absolute improvement of 18.3% and 7.4%, respectively, compared with no preprocessing and just face alignment.

Table 4. Result of different preprocessing.

| Methods               | Accuracy |
|-----------------------|----------|
| None                  | 62.8     |
| Face Alignment        | 73.7     |
| **Two-stage lip corrector(ours)** | **81.1** |

5. Conclusion
In this work, we use the proposed lip deflection classifier and the two-stage lip corrector to effectively solve the influence of large angle lip deflection on the accuracy of lip reading. The classification experiments prove that the ResNet18 is more sensitive to the tiny lip deflection. We also explored the combination of a variety of different network structures to further improve the quality of lip correction and the accuracy of lip reading. The lip corrector experiments show that StarGAN is suitable for translating the lip with different angle to the fixed angle in the first-stage, but the Pix2PixGAN is better in translating the fixed angle to fixed angle in the second-stage. Using the two-stage lip deflection corrector, we achieve an absolute improvement of 7.4% than the traditional face alignment methods, which will greatly increase the lip reading accuracy in the wild. We hope that the proposed lip corrector and deflection classifier could provide some valuable references to the related researchers.

References
[1] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks[C]// NIPS. Curran Associates Inc. 2012.
[2] Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[J], arXiv, 2017.
[3] Thompson, J.N. (1984) Insect Diversity and the Trophic Structure of Communities. In: Ecological Entomology. New York. pp. 165-178.
[4] Chung J S, Zisserman A. Lip Reading in the Wild[C]// Asian Conference on Computer Vision.
[4] Martinez B, Ma P, Petridis S, et al. Lipreading Using Temporal Convolutional Networks[J]. IEEE, 2020.
[5] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks[J]. Advances in Neural Information Processing Systems, 2014, 3:2672-2680.
[6] Szegedy C, Wei L, Jia Y , et al. Going Deeper with Convolutions[J]. IEEE Computer Society, 2014.
[7] Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2016.
[8] Martinez B, Ma P, Petridis S, et al. Lipreading Using Temporal Convolutional Networks[C]// ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
[9] Bai, S.; Kolter, J. Z.; and Koltun, V. 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. arXiv preprint arXiv:1803.01271.
[10] Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, "OuluVS2: A multi-view audio visual database for non-rigid mouth motion analysis," 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 2015, pp. 1-5
[11] Davis E King, “Dlib-ml: A machine learning toolkit,” The Journal of Machine Learning Research, vol. 10, pp. 1755– 1758, 2009.
[12] Isola P, Zhu J Y, Zhou T, et al. Image-to-Image Translation with Conditional Adversarial Networks[J]. IEEE, 2016.
[13] Zhu J Y, Park T, Isola P, et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks[J]. IEEE Computer Society, 2017.
[14] Kameoka H, Kaneko T, Tanaka K, et al. StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks[J]. IEEE, 2018.