Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 199 (2022) 1444–1449

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021)

A Comprehensive Dataset for Machine-Learning-based Lip-Reading Algorithm

Jin Ting, Chai Song*, Hongyang Huang, Taoling Tian

Southwest Minzu University
Chengdu, China
792174429@qq.com

Abstract

Lip-reading technology captures the content of the speaker by analyzing the characteristics of the mouth movement. It has a wide application prospect in the fields of daily life, security and so on. The training of the lip-reading model relies on a large amount of data, and the construction of the lip-reading dataset is the first step of lip-reading. The quality of the dataset greatly affects the work of the whole lip-reading system. Therefore, this paper carry out research on the construction of lip-reading dataset. First of all, frames are extracted from original videos by using the Scikit-Video. Then face detection is performed by applying dlib. Lip images are captured by processing the feature points to achieve lip cropping. Finally, data augmentation is performed to enlarged the dataset. The resulting dataset has 33 speakers, each with 7,000 pictures of their lips.

© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021)

Keywords—Lip-reading, face detection, lip cropping, dlib library, data augmentation

1. Introduction

Lip-reading is a technology based on machine vision and natural language processing, and has a wide range of application scenarios. For example, lip-reading can assist speech recognition. When the ambient noise is too high, the accuracy of speech recognition is affected because the audio is disturbed by the environment. Lip-reading only needs visual information, not audio information, so it can assist speech recognition to improve its accuracy. Lip-reading also is used to assist hearing impairment. There are about 20 million people with hearing and language disabilities in China, whose communication mainly rely on sign language. However, not everyone has mastered the sign language, lip-reading can make up for this shortcoming and help people communicate better.

The establishment of dataset is the primary problem of lip-reading, the success of subsequent lip feature

1877-0509 © 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021)
10.1016/j.procs.2022.01.183

extraction and lip-reading recognition are influenced by it. As the initial stage of lip-reading , a high quality dataset is the basic guarantee for the success of the lip-reading system. Therefore, the importance of the dataset is self-evident. The production of lip-reading dataset mainly has the following difficulties :(1) There are no faces in some parts of the collected video or faces occupy a small proportion of the picture, which affect the recognition accuracy if not processed; (2) The clipping range of the lip region also has an impact on the recognition result; (3) The training of the model requires a large number of samples, and the results of model training are affected when the number of samples collected is insufficient

In this paper, a dataset construction method is proposed. Firstly, the collected videos are processed by using the Scikit-Video library, and the processed videos are decomposed in the form of frames. Then the face detection model of dlib library is used to find the face from the image after video decomposition and extracting the required lip region. Finally, expanding the dataset by data augmentation.

The rest of this paper is organized as follow: Section 2 introduces the related work, the method is proposed in Section 3, Section 4 is the implementation and Section 5 concludes the paper.

2. Related Work

With the development of deep learning[1][2], more and more people begin to study lip-reading. University of Oxford[3], University of Washington[4-6] and many other famous universities, and Internet companies such as Google[7], vigorously carry out research in this field.. But compared with image processing, natural language processing and other fields, lip speech recognition is still in its infancy. The number of lip-recognition datasets is also smaller than in areas such as image processing.

In 2014, Kuniaki Noda[8] et al. from Waseda University used a variant of AlexNet[9] to extract features from lip images, combined with MFCC features of audio and GMM-HMM model, and finally achieved a word classification accuracy of 38% on a specific data set. In 2016, Chung&Zisserman[10] from VGG Group of University of Oxford published the LRW data set in the field of lip recognition. The data set contains 500 categories, and the VGG-M[11] model is used to model the image in the form of Multiple Tower, and the classification accuracy rate of 61.1% is achieved in LRW.

Mingmin Yang et al.[12]presents a public large-scale Mandarin lip-reading dataset named LRW-1000[9], which contains 1,000 classes with 718,018 samples from more than 2,000 individual speakers. The data set is based on videos from Chinese television shows. Tracking faces by Kernelized Correlation Filter(KCF), and to obtaining 80 key points of the face by using SeetaFaceEngine2. Then according to the location information, cutting out the required lip images. Randomly flipping images in the same sequence horizontally as an data augmentation step.

3. Proposed Method

The video used in this experiment is from the open source English lip-language data set GRID, which is recorded by 34 people (the video of record No. 21 is missing), and each person speaks 1000 sentences. The length of each video is about 3s and the frame rate is 25fps. After collecting the videos, the first step is using scikit-video to extract frames. Secondly, the dlib face detector is used to carry out face detection on the processed frame images. Discarding the images with no figures, the number of figures greater than one and the proportion of figures relatively small. Then lip detection was performed on the retained images, and lip images were cropped. Finally, the extracted lip image was enhanced to enlarge the data set. The overall design process is shown in Figure 1:

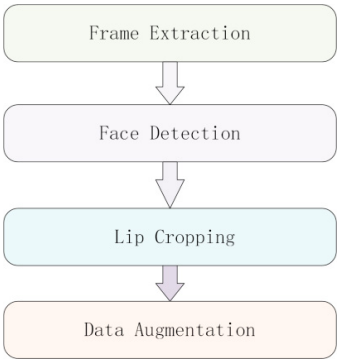

Fig. 1. Proposed Method

3.1. Frame Extraction

Scikit-video is a video processing library in Python that allows users to call various video processing algorithms. Skvideo.io is a video read/write module using FFmpeg/LibAV as the back end. Using the available back end, it will parse the video metadata with the appropriate probe (FFprobe, avprobe, or even mediaInfo). Here, the skvideo.io.FFmpegReader function is called to convert the video into a sequence of images for subsequent processing, as shown in Figure 2. And it returns the video shape in number of frames, height, width, and channels per pixel.

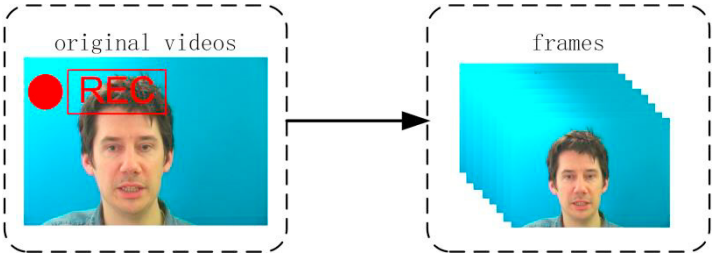

Fig. 2. Frame Extraction

3.2. Face Dection

Dlib is a cross-platform library written in C++ technology, which contains many algorithms for machine learning. It is very convenient to use. At present, dlib has been widely used in industry and academic fields, including embedded devices, robotics and high performance computing, etc. The steps of face detection using dlib are as follows: use the function dlib.get_frontal_face_detector to realize face detection. When there is no face in the frame, the frame is discarded, and the frame with only one face and a suitable proportion of face will be retained. The specific process is shown in Figure 3.

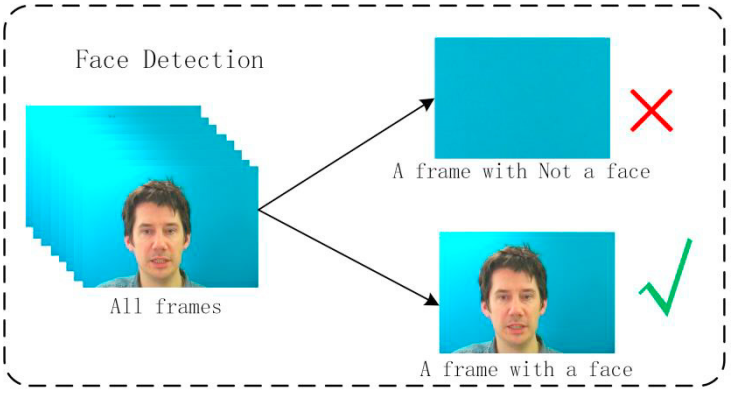

Fig. 3. Face Detection

3.3. Lip Cropping

The dlib library is used to detect 68 key points of the face, and the lip region needed was extracted from the marked key points and appropriately expanded. It is seen from the detected frames that the range of lips is 49 to 67 points. The main ideas are as follows: First, Select the left, right, top and bottom of the key points of the lips, which are defined as X_left , X_right , Y_left , Y_right ,to determine the range of lips; Secondly, the horizontal center point of the lip was calculated according to the left and right key points, and the vertical center point of the lip was calculated according to the top and bottom two key points. The expressions are as followed：

$$X\_center = (X\_left + X\_right)/2 \tag{1}$$

$$Y\_center = (Y\_left + Y\_right)/2 \tag{2}$$

Then determine the clipping range. Here, Set the extended distance to 15, that is, the range of the lip is centered on the calculated center point and expanded by 15 along the upper, lower, left and right sides respectively. Taking the left and right points as an example, *border* represents the expanded value. The calculation formulas are as follows:

$$X\_left\_new = X\_left - border \tag{3}$$

$$X\_right\_new = X\_right + border \tag{4}$$

Finally, cut the lips according to the expanded range to get the final lip picture.

3.4. Data Augmentation

The experiment shows that the more samples, the better the effect of the model and the stronger the generalization ability of the model. Data Augmentation can generate more equivalent (equally effective) data based on limited data, increasing the number and diversity of samples. After obtaining the required lip images, data enhancement can be used to reduce overfitting and improve the robustness of the model.There are many ways to enhance data, and I won't list them all. In this paper, the traditional image conversion series, such as brightening and darkening of the original lip image, horizontal mirroring and regular mirroring, Gaussian noise and so on, are used for data augmentation. Specific parameters are shown in Table 1.

Table 1. Specific parameters of data augmentation

No	methods	parameters
1	Brighten	Alpha = 1.5
2	Dim	Alpha = 0.5
3	Mirror	Horizontal mirror
4	Gasuss	Mean = 0,var = 0.008
5	Rotate_1	Rotation angle = -20°
6	Rotate_2	Rotation angle = +20°

4. Implementation

This project will conduct network model training on a 24GB NVIDIA Quadro P6000 GPU and a 56-thread Intel(R) Xeon(R) GOLD 5120 processor. The programming environment is Python.

The experimental results are shown in Figure 4. After one original frame is processed, seven different lip pictures can be obtained. 33 speakers' videos were produced in this data set and stored in 33 different folders. Each folder contains 5 trained models (1000 samples are divided into 5 parts), and each model contains data size of 200*7*75*60*120*3. 7 represents the original data clipped from the lips and the new data enhanced by six kinds of data augmentation, and 75 represents the video with a frame rate of 25fps and a length of 3S. 60*120 is the width and height of the video, and 3 represents the RGB type of image.

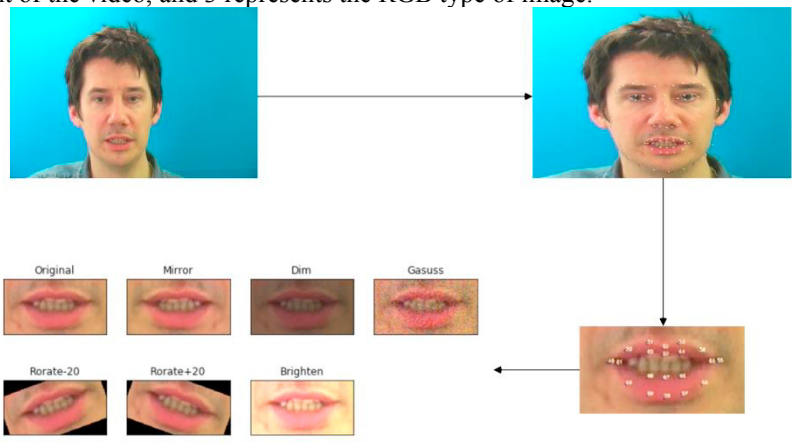

Fig.4. The results

5. Conclusion

In this paper, the problem of how to make a dataset in lip-reading is studied, and the overall design idea of dataset construction and the realization of each module are introduced. The experiment proves that the use of dlib and other libraries effectively and accurately identify the face and extract the lip region in the video. Finally, a high-quality dataset has been successfully built, which ensures the smooth development of the following steps in the process of lip-reading, such as feature extraction and lip reading recognition.

Acknowledgement

This paper is supported by the Key Research and Development Project of Sichuan Province (2021YFG0358), the Fundamental Research Funds for the Central Universities, Southwest Minzu University (2021PTJS24), and Southwest Minzu University Innovative Research Project of Graduate Students in 2021 (CX2021SZ39).

References

[1] Lecun Y , Bengio Y , Hinton G . Deep learning[J]. Nature, 2015, 521(7553):436-444.
[2] Hinton G E , Salakhutdinov R R . Reducing the Dimensionality of Data with Neural Networks[J]. Science, 313(5786):504-507.
[3] Chung, J. S. , et al. "Lip Reading Sentences in the Wild." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, 2017.
[4] Goldschen A J, Garcia O N, Petajan E. Continuous optical automatic speech recognition by lipreadin, Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on. IEEE, 1999:572-577 vol.1.
[5] Chiou G I, Hwang J N. Lipreading by Using Snakes, Principal Component Analysis, and Hidden Markov Models to Recognize Color Motion Video. IEEE Transactions on Image Processing, 1996.
[6] Chiou G I, Hwang J N. Lipreading from color video. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2002, 6(8):1192-1195.
[7] Shillingford B , Assael Y , Hoffman M W , et al. Large-Scale Visual Speech Recognition[J]. 2018.
[8] K. Noda, Y. Yamaguchi, K. Nakadai, et al. Lipreading using convolutional neural network. Made Available by the Northern Territory Library Via the Publications Act, 2014, 25(6): 840-849.
[9] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) Imagenet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems (NIPS), 1097-1105.
[10] Chung J S, Zisserman A. Lip reading in the wild. Asian Conference on Computer Vision. Springer, Cham, 2016: 87-103.
[11] Simonyan K , Zisserman A . Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.
[12] Yang S , Zhang Y , Feng D , et al. LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild[C]// 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE, 2019.