AUDIO-VISUAL ALIGNMENT MODEL WITH NEURAL NETWORK

Haiqin Yin
University of Rochester
hyin6@u.rochester.edu

Bo Wen
University of Rochester
bwen3@ur.rochester.edu

ABSTRACT

Synchronizing audio and video has always been a time-consuming and frustrating task for video editors. It often takes many hours and several times of re-editing back and forth to perfectly align the pictures and music. Thus, there exits a high demand for an automatic synchronization tool for professional and amateur video editors. The proposed system uses music track as a reference and then processes the video to match with the vocal through a convolutional neural network, which is trained using a single voice singing dataset URSing. The current model is not fully trained due to reasons of lacking of pre-training and small dataset. Modification and improvement will be conducted soon.

1. INTRODUCTION

Recent years, the advanced video editing software and powerful social media platforms such as Instagram, Snapchat and Tik Tok have lowered the requirements of becoming a video editor, everybody can become a vlogger, YouTuber or online celebrity. One of the most common yet popular content among video creators on these social media platforms is lip-syncing. However, it is rather difficult to synchronize the lip motion to music without noticeable inconsistency that would result in a degrade of the content quality. It usually takes many trials of re-filming or re-editing to adjust the one short video properly. Similarly, professional video editors for film and television shows also suffer from the same issue. The efforts they put into editing clips of different shooting angles or from different footage to a fixed audio track seems inevitable. This problem can be even more complicated when it comes to dealing live broadcasting transmission delay between audio and video. In this case, the development of an application or function in software that can automatically align the video clips to audio is motivated.

Some previous works have addressed the audiovisual synchronization problem from different perspectives. In 2017, Suwajanakorn has proposed a synthesis model that generate corresponding images of lip model based on audio input using a recurrent neural network [6]. Later in 2018, Eskimez proposed a face landmark model generation model using long short-term memory (LSTM) network, which give more flexibility to the video synthesis approach [2]. To deal with some occasions when the video is re-edited based on audio content, Helperin has proposed a different audiovisual alignment model that stretches or compress audio signal to match with the video clips [3]. The generation model is effective for dealing with fixed audio with video that has incorrect footage or missing footage of lip motion. However, the synthesis computation cost is higher than alignment when it comes to footage with correct video content falling on incorrect times. Our project therefore aims to build a audiovisual alignment model with convolutive neural network (CNN) that can detect and relink correct video clips based on fixed vocal track of a song to generate a video that would match with the music.

The paper is organized as follows. Section 2 describe some previous models that also aim for realizing audio-visual alignment and introduce the preprocessing methods used and the architecture of the CNN. Section 3 contains the information regarding the dataset we used and details of the training, testing and evaluation of the CNN model. In Section 4, we discuss the result of the evaluation and reach a conclusion of how well our proposed method works. In the end, the potential reasons for conclusion and future direction of the development of the method are presented.

2. METHOD

Sliding window and regression are two current deep learning approaches that are used for audio-visual alignment problems. Sliding window approach calculates the correlation values of each audio frame in an audio segment with a sliding window and uses the frame number with highest correlation for correspondence [4]. Regression approach, in contrast, predicts the correct frame position only by looking at the corresponding audio content, which is proven to be faster than the sliding window approach [5].

Different than the two methods mentioned above, our approach uses binary classification to find the corresponding relationship between an audio clip and a video frame. A convolutional neural network (CNN), which takes extracted features of segmented audio content and video image through the preprocessing stage, is used to train a model that classifies whether a pair of preprocessed audio and video clip match each other or not. The output audio and video temporal sequence will be evaluated using dynamic time warping (DTW), which is an algorithms used to measure the similarity, are used as a guideline to adjust the time position of video frames.

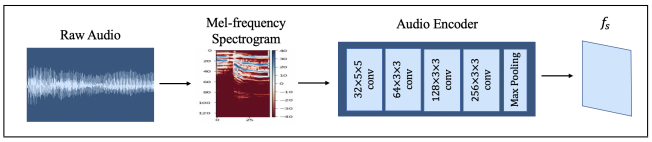

Figure 1. Audio feature extraction

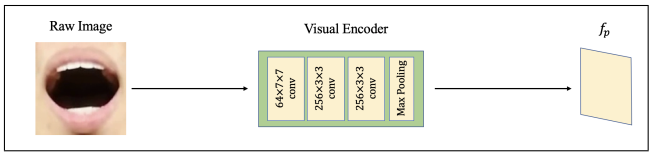

Figure 2. Video feature extraction

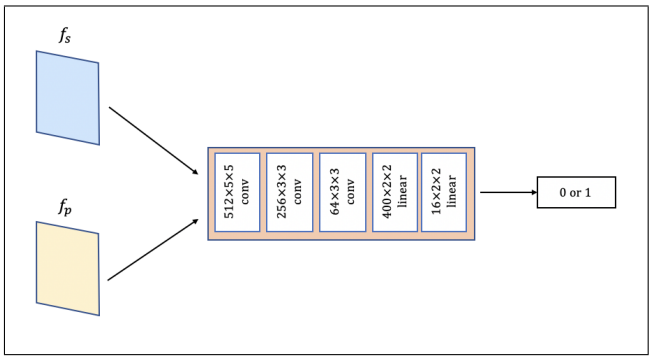

Figure 3. CNN Architecture

2.1 Preprocessing

The audio and video are preprocessed separately before being fed into the neural network. The preprocessing method is adapted from [1]. The purpose of the preprocessing stage was to get an abstract representation of the audio features and video features. The details are presented as follow.

2.1.1 Audio feature

As the audio and video are considered as a pair of inputs, and the video clips is 16 frames long, the audio of the song is correspondingly segmented into snippets of 0.64 seconds and then transformed into the time-frequency (T-F) domain by acquiring its Log-amplitude Mel-frequency Spectrogram (LMS) with window length of 1024, hop size of 512 and Mel-bands of 128. Log-amplitude amplitude Mel-spectrogram provides a time and frequency representation incorporating human perception. It outperforms other audio feature representation such as regular Mel-spectrogram, STFT and MFCC in music information retrieval. The LMS is processed through an audio encoder, which consists of 4 layers of convolution and 1 layer of max pooling to extract audio features $f_s \in R^{T\times F}$ [1]. T and F refer to the number of time frames and frequency channels.

2.1.2 Visual feature

The video in the dataset we used is filmed at a frame rate of 25 FPS. The video is segmented into clips of 16 frames as a single lip model is found to typically lasts for that long. Since the mouth area is assumed to be more correlated to the audio than other area of face, the video is cropped with a resolution of 64 $\times$ 64 pixels centering at the mouth. The image extracted from the video frame is then processed through a visual encoder, which consists of 3 convolution layer and 1 max pooling to get the image feature $f_p \in R^{H\times W}$ [1]. H and W refer to the height and width of the image, which are both 64 pixels in this case.

2.2 CNN Architecture

The audio time-frequency feature f_s and image spatial feature f_p acquired from the preprocessing step are then concatenated to create a spatio-temporal feature. The spatio-temporal feature are combined and forward into a tensor in the proposed CNN model, which consists of three convolutional layers and two linear layers that uses a sigmoid function. The dimension details of layers in the CNN architecture is shown in Fig.3. Through the network, an output of a binary number of 1 or 0 can be generated and represent whether the two inputs match or not.

3. EXPERIMENT

3.1 Dataset

We recorded the audio-visual solo singing performance dataset URSing and used for training and testing the neural network in this project. The dataset contains 65 single-voice English and Mandarin Chinese songs in different genres and was recorded by a mixture of male and female. The vocal tracks were recorded in 44.1kHz and 16bit mono and then convert to stereo in the mixing process. The recording is conducted in isolation environment and mixed using noise reduction, compression and reverb effects. Clean vocals and instrumental are also separately provided and therefore only the clean vocals are used for training and the source separation step, which is required in application, is omitted.

As the audio and video of a singing sample in the dataset we used is intrinsically matched, parts of the original data are altered to include some unmatched samples and thus increase the diversity of the data. The `excerpt_bv_rev` folder, which contains 26 video clips that all lasts for 30 seconds, was used as the primary dataset. In the 26 audio and video clip pairs, 13 of them are randomly selected for training, 3 clips are for validation and 10 clips are for testing. A approximate ratio of 5:1:4 is maintained.

3.2 Training

Each of the 13 audio and video clips in this sample folder was firstly edited and segmented into 46 files as stated in the prepossessing section accordingly and resulted in 598 files in total. Then the data stream of audio and video clips, together with a corresponding label list, which holds the ground truths of match of unlatch, are written into a new pickle file that converts the list into a character stream. In the next step, an audio excerpt and a video excerpt are randomly selected and a label is randomly generated. The

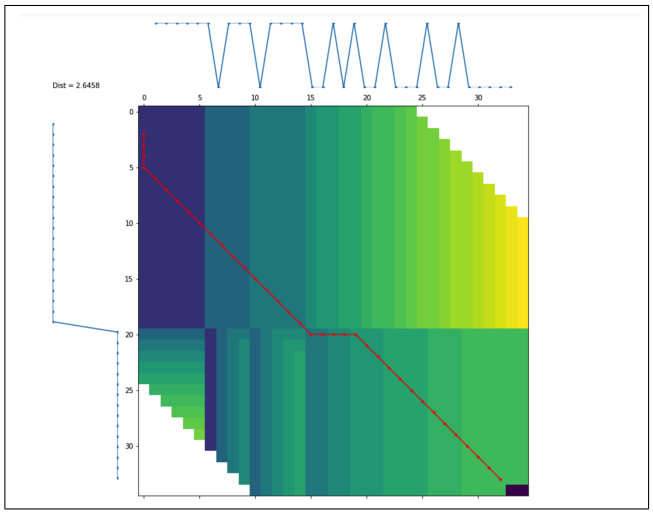

Figure 4. DTW with network output list (horizontal) and ground truth label list (vertical)

name of the video file is renamed to match the audio file give the condition that the randomly generated label is 1. Otherwise if the label is 0, the audio file and video file will be re-selected until they are not matched. The purpose of this operation is to make sure the ground truth is not necessarily be all true. Another three pairs of clips randomly selected from the rest of the samples are used to validate if the model works.

3.3 Test and Evaluation

Once the validation step proves the executability of the model, the 10 testing audio and video pairs are used to test the performance of the network. Evaluation is conducted by counting the number of label generated that match the ground truth label. An accuracy rate is calculated to intuitively show the performance of the proposed network.

4. RESULT

The result of the current version of the network shows that the model is not fully trained and thus further modification and improvement is required. The potential reason are lack of pre-training and in-flowing data.

5. CONCLUSION

For future improvement, we will be focusing on modifying the network and preprocessing method to gain a valid result. Post-processing of the video based on the result generated from the neural network. The DTW generated using the result of the neural network can be used as guideline to shift the frames accordingly and match with the true audio content. In addition, the network proposed can only deal with unmatched video without missing and false lip motion. A modification of our model should be investigated to deal with such situation.

6. REFERENCES

[1] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018.

[2] Sefik Emre Eskimez, Ross K Maddox, Chenliang Xu, and Zhiyao Duan. Generating Talking Face Landmarks from Speech. In Yannick Deville, Sharon Gannot, Russell Mason, Mark D Plumbley, and Dominic Ward, editors, *Latent Variable Analysis and Signal Separation*, pages 372–381, Cham, 2018. Springer International Publishing.

[3] Tavi Halperin, Ariel Ephrat, and Shmuel Peleg. Dynamic temporal alignment of speech to lips. *CoRR*, abs/1808.06250, 2018.

[4] Yuyu Liu and Yoichi Sato. Recovery of audio-to-video synchronization through analysis of cross-modality correlation. *Pattern Recognition Letters*, 31(8):696–701, 2010.

[5] Toshiki Kikuchi; Yuko Ozasa. Watch, listen once, and sync: Audio-visual synchronization with multi-modal regression cnn. 2018.

[6] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4):95:1–95:13, July 2017.