Audiovisual Alignment Model with Neural Network

Bo Wen & Haiqin Yin

Dept. of Electrical and Computer Engineering, University of Rochester

ABSTRACT

Synchronizing audio and video has always been a time-consuming and frustrating work for video editors. It often takes a several time of re-editing to perfectly align the pictures and music. Thus, a high demand of an automatic synchronization tool for professional and amateur video editors exists. The proposed system uses music track as reference and process the video to match with the vocal through a convolutional neural network, which is trained using a single voice singing dataset URSing.

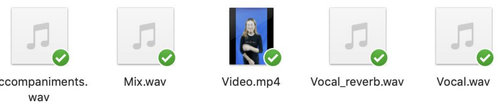
METHODS

The proposed audiovisual alignment model 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.

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 songs
  - English and Mandarin Chinese
  - Different genres
  - Recorded by a mixture of male and female
  - 44.1kHz and 16bit mono
  - Converted to stereo in mixing
  - Recorded in isolation environment
  - Mixed using noise reduction, compression and reverb effects.
  - Clean vocals and instrumental are also separately provided .
- *A subset of URSing - excerpt_bv_rev*
  - *26 video & audio clip pairs*
  - *Duration: 30 sec*
  - *13 for training*
  - *3 for validation*
  - *10 for testing*



PREPROCESSING

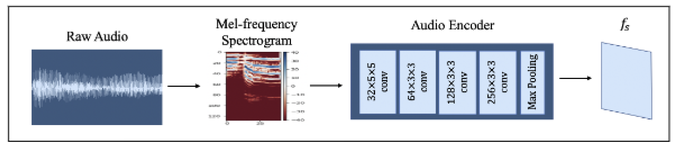


Figure 1. Audio feature extraction

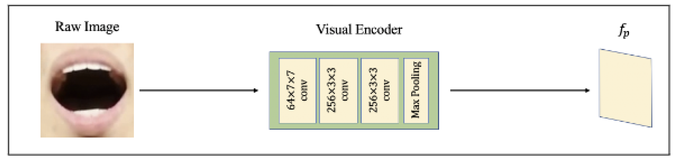


Figure 2. Video feature extraction

Audio Feature:

- Log-scaled Mel-frequency Spectrogram

Visual Feature:

- Segmented into 16 frames
- Cropped with resolution of 64× 64 pixels around mouth

CNN ARCHITECTURE

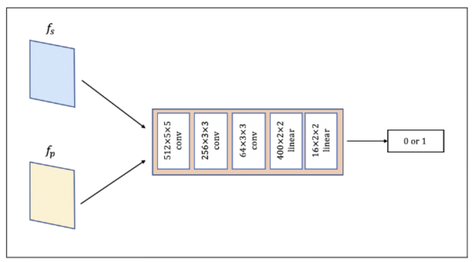


Figure 3. CNN Architecture

The audio time-frequency feature fs and image spatial feature fp 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 one linear layer 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.

RESULT & CONCLUSION

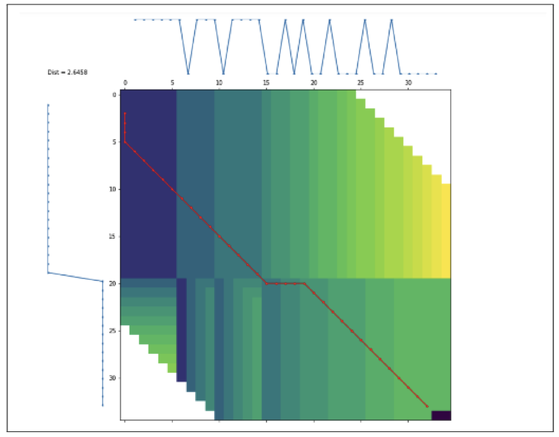


Figure 4. DTW with network output list (horizontal) and ground truth label list (vertical)

- Current result does not show that the network is fully trained
- Potential reasons might be lack of pre-training or the low learning efficiency based on the current preprocessing method
- Further modification of the network is needed

FUTURE WORK

For future improvement, we will be focusing on processing 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.

REFERENCE

[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.