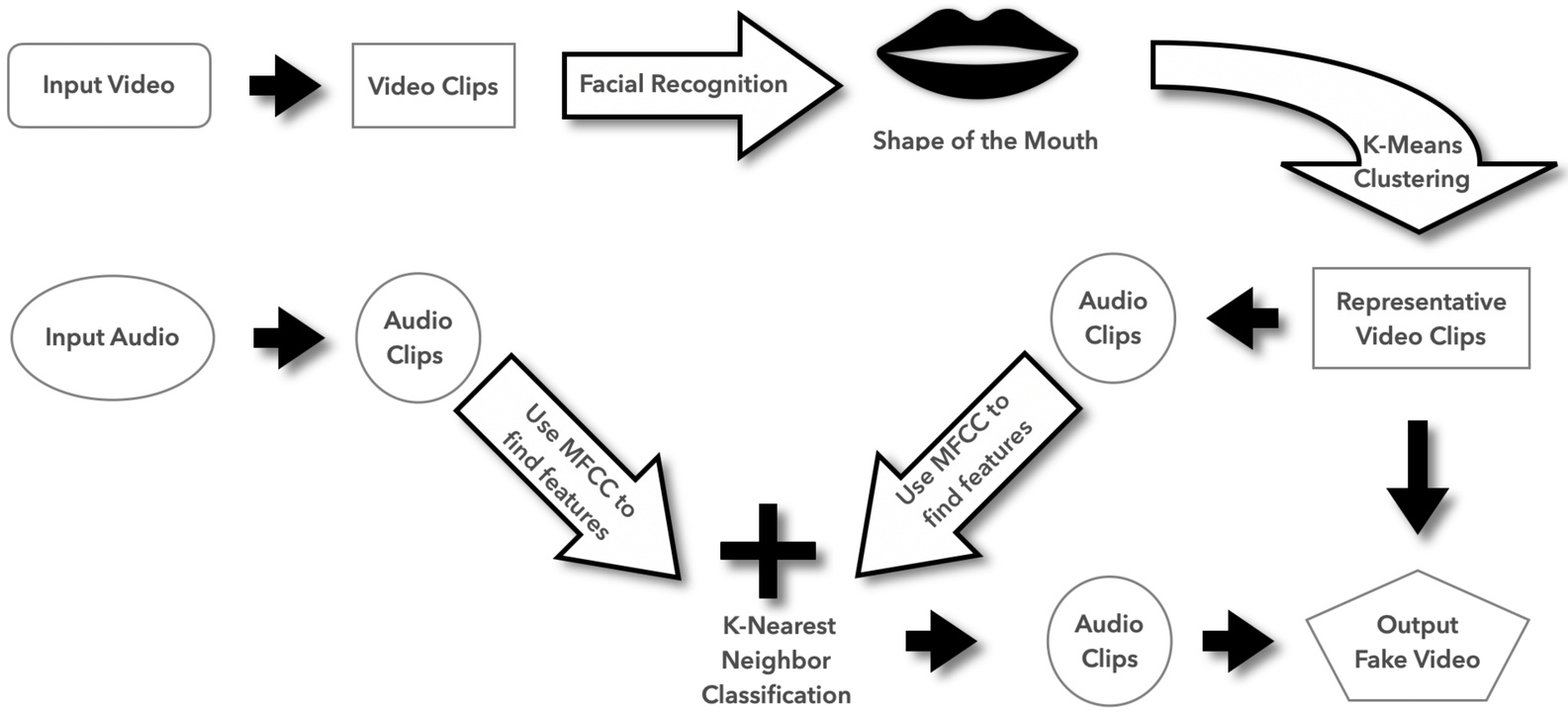
Fake Talking Video – Mapping from audio to video.

Weijie Lyu, Suyi Jiang, Zhiyuan Liang, Zhuyun Yin, Zhouyang Fang

1. Introduction

With the development of machine learning techniques and the increasing amount of data, computer vision-related tasks are more pervasive not only in laboratories and industries but also in our daily lives. Fake talking video is an exciting computer vision application. Existing methods often take advantage of the complicated state-of-art GANs with long-time training. Though incredibly plausible the imitation is, many computer vision enthusiasts cannot try it out on their own devices due to its heavy training cost. Herein we present a new method, generating fake videos from a series of video clips determined by any piece of audio using the K-Nearest Neighbor algorithm. To accelerate the speed of this algorithm and achieve better performance, we use K-mean clustering to grouping the video clips based on the shape of the mouth to get a standard mouth shape and its corresponding video clip to a specific audio clip. We show that given audio about one minute, we can output a fake talking video within less than 30 seconds. Besides, the methods we use are straightforward to understand and the training process is extremely fast.

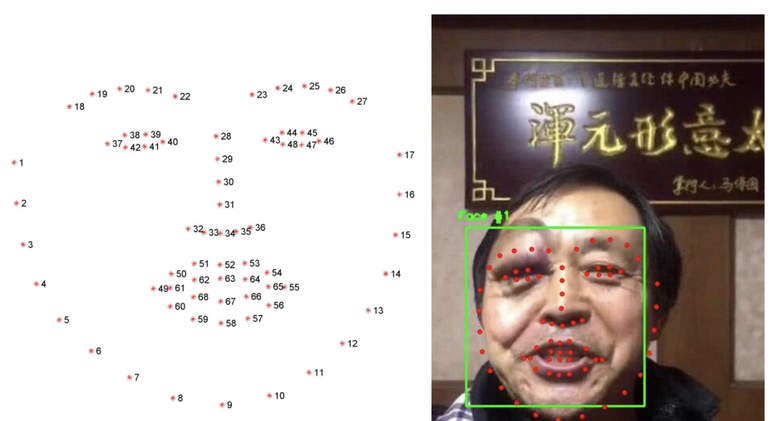
2.Flow Diagram



3. Methods

1 Mouth Shape Detection

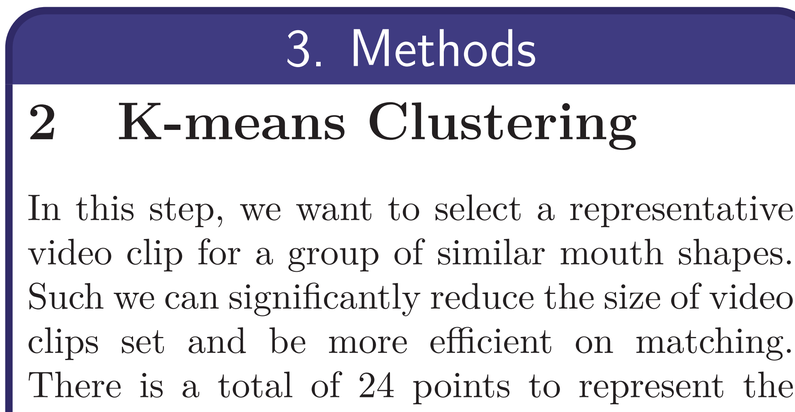
For the purpose of audio-video matching, we want to first construct a video-clips dictionary to store each kind of audio and its corresponding mouth movement. An input video will be slicing into video clips of different lengths, which range from 0.2s to 0.8s. For each video clip, we used facial feature detection methods to detect the shape of the mouth and used K-means clustering to build a code book of the mouth shapes and their corresponding audio clips.



3. Methods

2 K-means Clustering

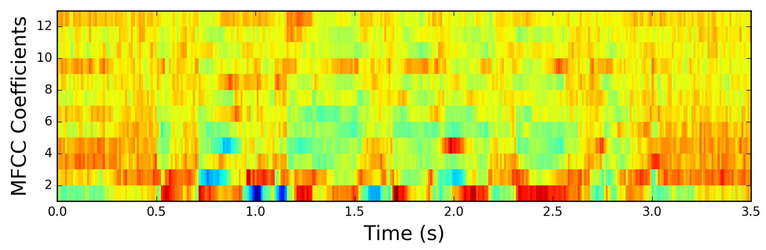
In this step, we want to select a representative video clip for a group of similar mouth shapes. Such we can significantly reduce the size of video clips set and be more efficient on matching. There is a total of 24 points to represent the position and shape of the mouth, 12 on the upper lip and 12 on the lower lip. We calculate the distance between the upper lip and lower lip by subtracting corresponding points then divide the length of the mouth and get a 12*2 matrix. We do K-means clustering base on these 12*2 matrices and get the centers as our video clip dictionary.



3. Methods

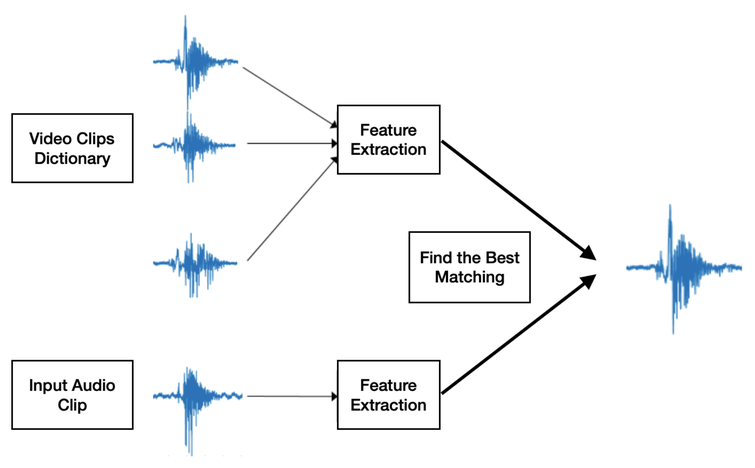
3 Audio Feature Extraction

In sound processing, the Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip. We extract the audio features based on their MFCCs.



4 KNN Matching

The input audio will first be sliding into audio clips of 7 different lengths. We use these audio clips to find the best match video clips from our video-clips dictionary respectively, and we choose the video clip with the highest audio matching score. Notice that the longer clips will always have a larger matching distance compared to the shorter clips, so when we compare their matching result, we divide the matching distance by the dimension of the clip to get a matching score.



4.Result

With mouth futures detection and corresponding K-means clustering, video clips can have a one-to-one match with any audio clips, thus form a plausible video we desired. Plus, the most exciting thing is once we spend some time clustering video clips, then various fake videos can be produced after feeding any audio almost in no time.