CSCI 5561 2019 Fall Project Proposal: Dubbing for Soundless Cat Video

Wei Dai
University of Minnesota, Twin Cities
Minneapolis, MN
dai00074@umn.edu

Chen Hu
University of Minnesota, Twin Cities
Minneapolis, MN
huxxx853@umn.edu

Abstract

In this project, we plan to do an AI dubbing for soundless videos of cats. We will be able to tell when cats would miaow from their mouth movement. In advance, we will tell what type of sounds they would generated by recognizing cats' facial expression and relating those to cats' emotion.

1. Introduction

In real word, vision changes are always accompanied by sounds. By deeply exploring the natural relation between vision and sound, scientists can make artificial intelligence better understanding the world. Researchers have done a lot of phenomenal studies in this field. Andrew Owens et al. conducted their research on predicting generated sound based on sound-waived videos in which stick hit certain materials. They claimed that if their model was well-defined, then computer vision techniques can help humans to analyze the physical properties of materials [7]. In advance, Yipin Zhou et al.'s research predicted raw audio signals from soundless videos[12]. Hang Zhao et al., in a similar way, tried to find relation between motion and sound by tracking players' hand motion and music their played[11].

Inspired by these researches, we plan to do an AI dubbing for soundless videos in this project. Due to our lack of research experience and domain knowledge in sound signal, we simplify the task as a video action/activity recognition and facial recognition task, instead of sound wave generation from soundless video. We want to apply the task into a small and fun area: cat video. The ultimate goal of this project is to interpret cats' emotion by recognizing their facial expressions, and then play corresponding sounds, such as angry sound or happy sound. The detailed method and work flow will be shown in Section 3. To emphasis,the crucial part of this project is the emotion detection of cats, and our major task would be improving the algorithm of emotion detection such that it can be better than current research results, which would be shown in preliminary result.

In the rest of the proposal, Section 2 will summarize related papers, and Section 3 will describe the picked baseline method and demonstrate our dubbing solution for soundless video of cats.

2. Related Work

Our project closely relates to video action detection and animal emotion classification tasks.

Since the application in this project is to dub for soundless cat video, an algorithm should be applied to detect whether a cat opened its mouth. This task generally belongs to video action recognition tasks. Jeff Donahue et al. used long-term recurrent convolution networks to extract video frame features and learn the relationships between sequential frames. Then, the model can make classification [2]. Wenbo Li et al. further utilized recursive neural network to build a RNN Tree for large-scale human action recognition [5]. Some closer studies include eye blink detection or driver drowsiness detection. Taner Danisman et al. proposed an old fashion method in computer vision to detect eye blink by calculating horizontal symmetry of eyes [1]. Some newer methods on the same task use deep neural network to detect eyeblink and drowsiness. Bhargava Reddy et al. demonstrated an effective deep learning method to detect drowsiness, and further compressed the model to fit real time detection [9].

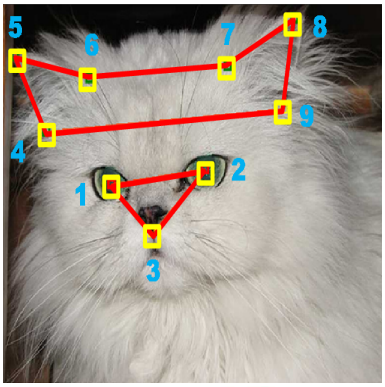
In our project, after detected whether a cat is meowing, we will recognize what kind of sounds should be played for the cat. Human facial expression detection is kind of related to this task. Some studies use local patches such as eyes, mouth and nose, as key features to train the model. Yingruo Fan et al. adapted a multi-region ensemble convolutional neural network to recognize facial expression [3] More human facial expression recognition related research can be found in [4]. A closer related work is about sheep emotion detection. Marwa Mahmoud et al. made their own labeled sheep emotion dataset and regarded sheep nose, mouse and ears patches as key features to classify sheep emotions by SVM method [6].

3. Preliminary Result

In this project, we intend to apply video action detection techniques to a novel application area: dubbing for cat video when cat meowing. Additionally, we intend to use more advanced methods such as CNN to solve the animal emotion detection problem. We choose method in [6] as the baseline in our project, and further apply our method into dubbing project.

3.1. Dataset

In order to adapt eye blink detection technique to realize the cat mouth open detection, we firstly need detect cat head in the video frames. We will take advantage of the cat annotation dataset [10].



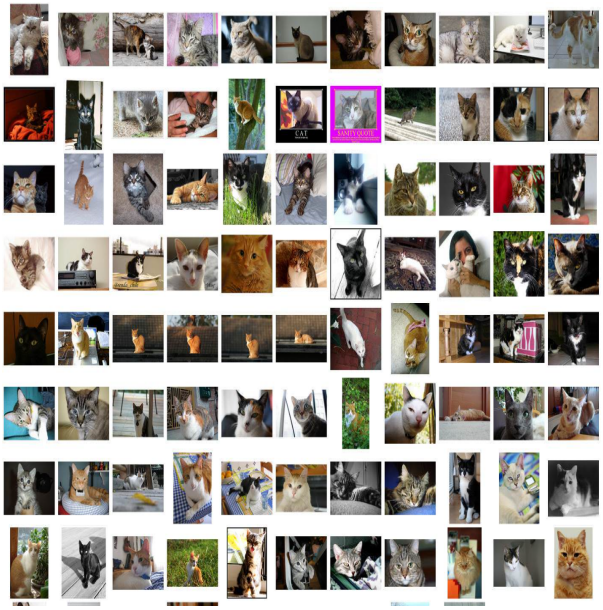


Table 1. Cat Annotation Dataset

Despite the lack of formal dataset of cats or kitties' emotion, there are thousands of cat video on YouTube. We might generate a small labeled cat emotion and action dataset from the videos by ourselves. Fortunately, we found a labeled cat sound dataset which includes 10 types of cat sounds[8].We might not use all types of the sound but it gives us a good reference to select matched cat voice.

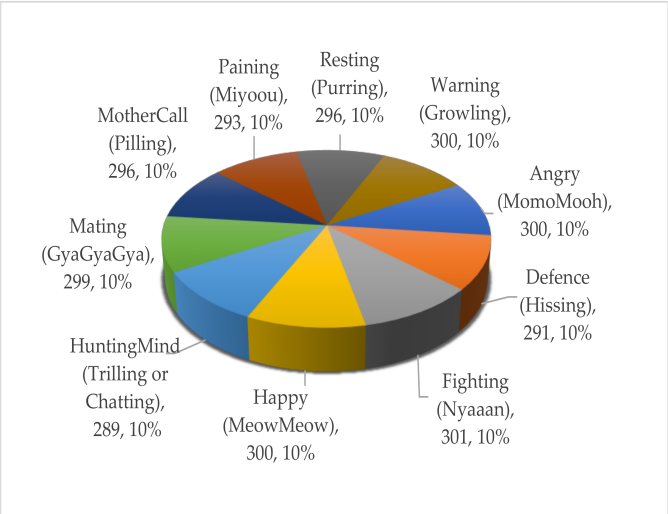


Table 2. Labeled Cat Sound Dataset

3.2. Preliminary Baseline Result

Due to the lack of cat data, currently we do not have the result of baseline method on cat data. In this subsection, we show the method and result in [6]. In the original paper, the sheep face detection pipeline is (a) initial face detection, (b) central landmark detection using Dlib, (c) full set of landmarks detected, then (d) face normalization, and their final average accuracy reached 0.67.

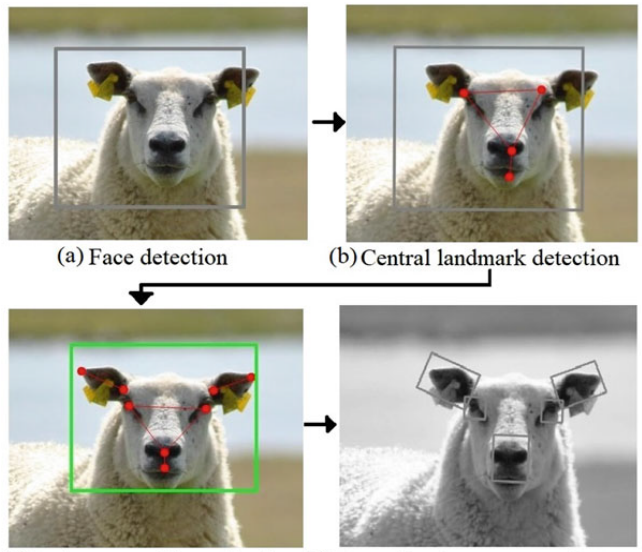


Table 3. Baseline Method

3.3. Proposed Method and Workflow Diagram

To realize our final goal, we can break our project into three related parts: cat face detection, cat video mouth open detection and cat emotion detection. Cat face detection will

help us locate cat and extract key feature patches like mouth, eyes and ears. Cat video mouth open detection will help us decide when to play a matched cat voice. Finally, Cat emotion detection will help us recognize the emotion of the cat in video and we can further choose a proper cat voice. In this project, we will focus on the improvement of the third part. The other two parts we will consider applying the state of art method to reach the goal. The three parts will work as the following diagram.

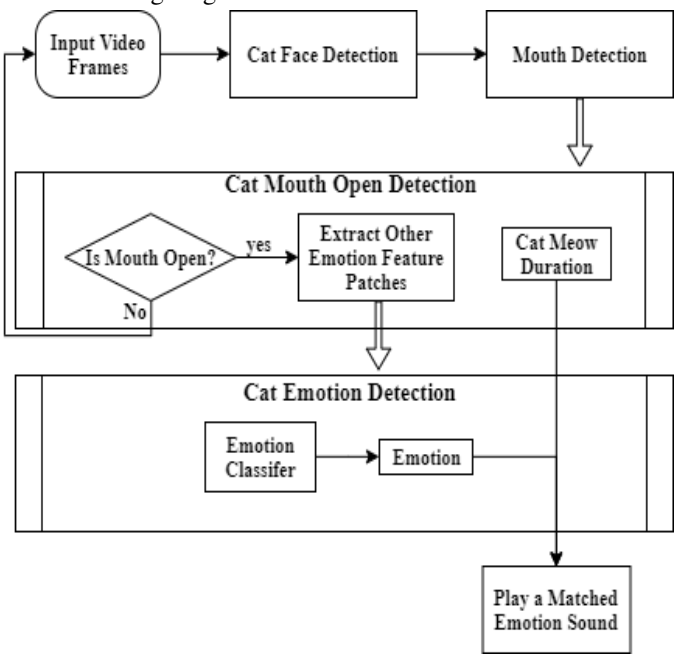


Table 4. Our Project Workflow

References

[1] T. Danisman, I. M. Bilasco, C. Djeraba, and N. Ihaddadene. Drowsy driver detection system using eye blink patterns. In *2010 International Conference on Machine and Web Intelligence*, pages 230–233, Oct 2010.

[2] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[3] Yingruo Fan, Jacqueline C. K. Lam, and Victor O. K. Li. Multi-region ensemble convolutional neural network for facial expression recognition. In Věra Kůrková, Yannis Manolopoulos, Barbara Hammer, Lazaros Iliadis, and Ilias Maglogiannis, editors, *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 84–94, Cham, 2018. Springer International Publishing.

[4] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *CoRR*, abs/1804.08348, 2018.

[5] Wenbo Li, Longyin Wen, Ming-Ching Chang, Ser Nam Lim, and Siwei Lyu. Adaptive rnn tree for large-scale human action recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[6] Marwa Mahmoud, Yiting Lu, Xijie Hou, Krista McLennan, and Peter Robinson. *Estimation of Pain in Sheep Using Computer Vision*, pages 145–157. Springer International Publishing, Cham, 2018.

[7] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. Visually indicated sounds. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] Yagya Raj Pandeya, Dongwhoon Kim, and Joonwhoan Lee. Domestic cat sound classification using learned features from deep neural nets. *Applied Sciences*, 8(10), 2018.

[9] Bhargava Reddy, Ye-Hoon Kim, Sojung Yun, Chanwon Seo, and Junik Jang. Real-time driver drowsiness detection for embedded system using model compression of deep neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[10] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat annotation dataset merged. 2008.

[11] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. *CoRR*, abs/1904.05979, 2019.

[12] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L. Berg. Visual to sound: Generating natural sound for videos in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.