# Emotion Recognition Based 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

*Emotion recognition has been well studied for human, but not for animals. In this paper, we will present a study on emotion recognition for cats. Finally, we would show an AI dubbing for soundless videos of cats. We will 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' emotions.*

## 1. Introduction

Facial expression is one of the most powerful, natural signals people use to express emotions. Lots of researches have been done in this field. The recent survey [4] showed an evolution of facial expression in terms of datasets and methods. However, few researches have been done by applying those methods to animals. In 2018, a paper recognized sheep level of pain by using classical HOG and SVM method [6]. Also, as pets are more and more adopted as family member, there is an increase demand for pets owners to understand what their pets are feeling. However, while pets can become stress reliever for owners, owners sometimes cannot really recognize pets' emotions. We want to build the communication bridge for pets and their owners. In this paper, advanced deep learning and computer vision techniques will be applied to recognize cat emotions.

Cat facial expression can be hard to justify. We would like to use a second media, cat sounds as a justifier as our work. In real word, vision changes are always accompanied by sounds and can be correlated with sound. For example, 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[14]. 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[13]. Sound and motion generated by live animals would also have relation. Maowing sound and cat motion would reflect the cat emotion/mood in current state. For example, if a cat was petted by its own, its facial expression should be relaxed and should generate cozy sound.

Inspired by these researches, we will present an AI dubbing for soundless cat videos in this paper. We simplify the task as a video action/activity recognition and facial recognition task. 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 generated sound can be compared with the original sound to show its correctness. 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 recognition such that it can be better than current research results, which would be shown in baseline method.

In the rest of the proposal, Section 2 will summarize related papers, and Section 3 will describe the picked baseline method and our proposed method. Section 4 will demonstrate our cat face detection, feature extraction and emotion recognition results together with dubbing solution for soundless video of cats. Section 5 will present potential future work and section 6 will be a brief summary.

## 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. Method

In this section, we will first present our chosen baseline method, and then we will state our proposed method in detail. We will also show our datesets in the last sub-section.

### 3.1. Baseline Method

We choose the method and result in [6] as our baseline, since it's most related to our task– animal emotion recognition. 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. For part (a) in pipeline, they used Viola-Jones object detection framework [11] to implement the frontal face detection, which is introduced in the lecture. In landmark detection, they used HOG descriptor to extract features and then SVM as classifier. For their result, their final average accuracy reached 0.67 in sheep dataset. As we can see, even this paper was published in 2018, they didn't use modern deep learning techniques. Even facial expression recognition for human is a well-known problem in computer vision community, it hasn't been widely applied to other areas. We would like to apply modern techniques to this problem and prove it will outcome better performance. Here below is our proposed method.

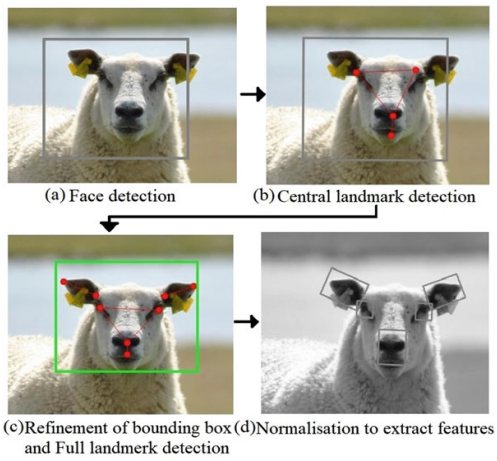



Figure 1. Baseline Method

### 3.2. Proposed Method

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 recognition. 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 in Fig. 2. Detailed methods of each part are listed in the following subsections.

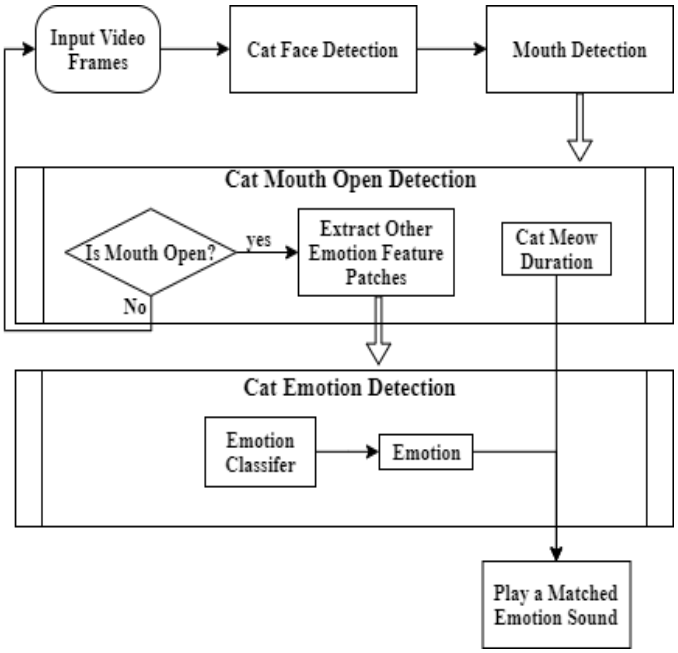



Figure 2. Our Project Workflow

#### 3.2.1 Cat Face Detection

Deep neural network-based face recognition techniques and object recognition is relative mature in computer vision community. We adapted one of modern frameworks, Darknet Yolov3 [10], to detect cat face in the video frame images.

#### 3.2.2 Cat Mouth Open/Close Detection

We transferred the mouth open/close detection problem into an image binary classification problem. To reach a higher performance and accelerate the training, we adapted the idea of transfer learning. We first used the feature extraction layers of a pretrained VGG16 net as a feature extractor to extract features of cat face images. Then, a classifier using fully connected layer was trained. Finally, plug pretrained VGG network architecture back and fine-tuned the weights of last or last two convolution layers.

#### 3.2.3 Cat Emotion Recognition

Due to limited time, we only consider three common emotion of cat, happy, angry, and sad. Similarly, we transferred emotion recognition problem into multi-class image classification problem. We tried two strategies to solve the problem. One is taking the whole cat face as input image and doing classification. The other is taking images of eyes, ears, mouth as input separately as shown in Fig 3. The rest of process is similar to cat mouth open/close detection.

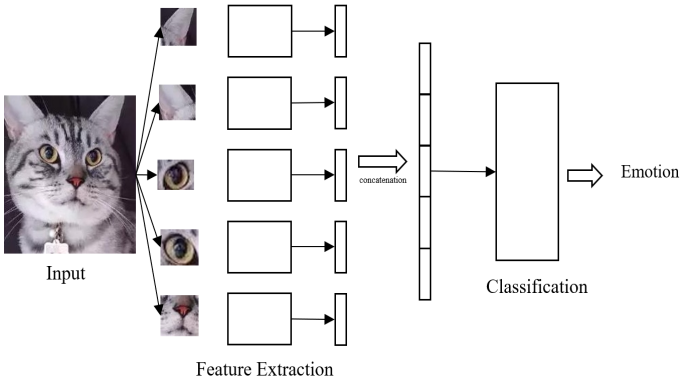



Figure 3. Cat Emotion Detection Strategy 2

### 3.3. Dataset

We mainly utilized four types of datasets to finish our work. They are for cat face detection, cat mouth open/close detection, cat emotion recognition, and cat sound play respectively.

#### 3.3.1 Cat Face Detection

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 [12]. This dataset includes annotated cat eyes, ears, mouse points. We further constructed bounding box for each position by inducting from the geometry relationship between ears, eyes, mouth, and face .

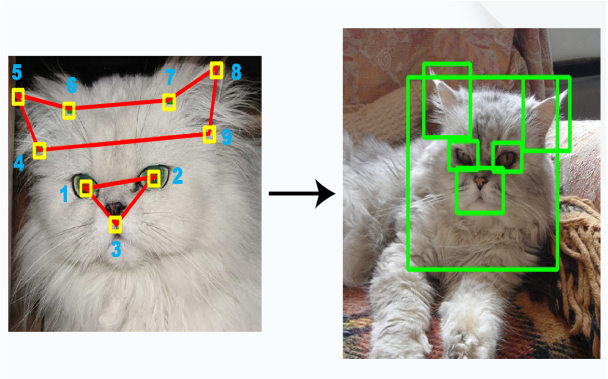



Figure 4. Cat Annotation Dataset

#### 3.3.2 Cat Emotion Data

We collect cat emotion data ourselves from both Chinese and English social medias, such as YouTube and Bilibili. We get the labels of emotion by searching keywords like angry cats or happy cats.

#### 3.3.3 Cat Mouth Open/Close Data

We collect cat emotion data ourselves from cats' meow video. It is straightforward to distinguish the frame images from mouth close and open.

#### 3.3.4 Cat Sound Data

To finish our ultimate application, dubbing for soundless cat video, a cat emotion sound dataset was adapted. It contains 10 types of labeled cat sounds[8]. It gives us a good reference to select matched cat voice. We will use three types of them.

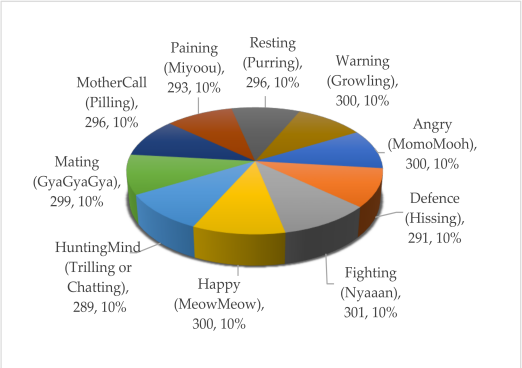



Figure 5. Labeled Cat Sound Dataset

## 4. Results

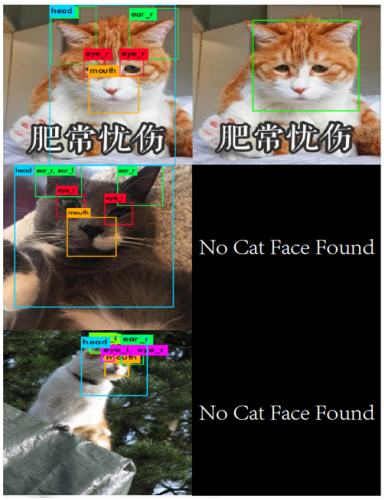
To be consistent with the base line method, we will compare the quantitative and qualitative results for cat face detection and cat emotion detection here and list quantitative and qualitative result for cat mouth open/close detection.

### 4.1. Cat Face Detection

In our project, we use Darknet Yolov3 framework while the baseline method used Viola-Jones object detection framework to detect cat face and facial landmark. The result shows that Yolov3 is more robust to the color of cat and rotation of cat head. The detection accuracy and MAP is reported in table below and the visual comparison is shown in Fig. 6.

| Method | MAP | Accuracy |
|---|---|---|
| Baseline | 0.5862 | 0.2789 |
| Yolov3 | 0.7345 | 0.9055 |

Table 1. Cat Face Detection Result. Yolov3 is better.




Yolov3 Baseline

Figure 6. Cat Face Detection Result

### 4.2. Cat Mouth Open/Close Detection

We detect cat mouth open/close frame by frame. The visual result is shown as Fig. 8, classification result is reported in table below, and confusion matrix is shown in Fig. 7.

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| Close | 0.97 | 0.92 | 0.94 |
| Open | 0.86 | 0.95 | 0.91 |

Table 2. Cat Mouth Open/Close Detection Classification Report

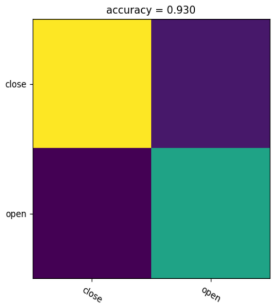



Figure 7. Cat Mouth Open/Close Detection Confusion Matrix

### 4.3. Cat Emotion Detection

The baseline method utilized HOG descriptor to extract animal eyes, ears, and mouth features and use SVM as classifier. In this project, we adapted modern neural network framework and tried different methods to recognize cat emotion. The classification results are shown as table and confusion matrices below. Due to the limited size of dataset, the total number of test image is 120 where 40 for each class.

| Method | Class | Precision | Recall | F1-score |
|---|---|---|---|---|
| Baseline | Sad | 0 | 0 | 0 |
| Our_head | Sad | 0.49 | 0.60 | 0.54 |
| Our_landmark | Sad | **0.72** | **0.78** | **0.75** |
| Baseline | Happy | 0 | 0 | 0 |
| Our_head | Happy | 0.60 | 0.36 | 0.45 |
| Our_landmark | Happy | **0.82** | **0.57** | **0.68** |
| Baseline | Angry | **1.00** | 0.33 | 0.50 |
| Our_head | Angry | 0.59 | 0.70 | 0.64 |
| Our_landmark | Angry | 0.65 | **0.80** | **0.72** |

Table 3. Cat Emotion Detection Classification Report

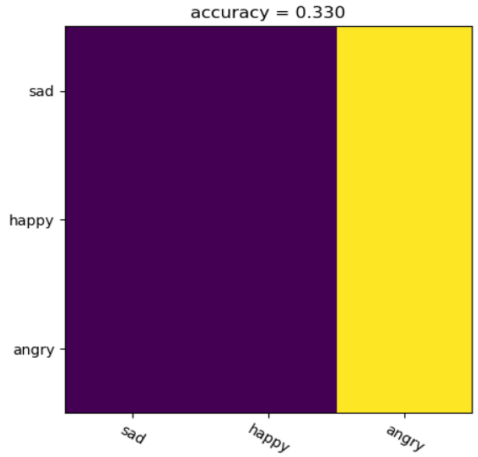



Figure 9. Cat Emotion Detection Result: Baseline Accuracy and Confusion Matrix

Figure 8. Cat Mouth Open/Close Result: Green box indicates closed mouth and red box indicates opened mouth.

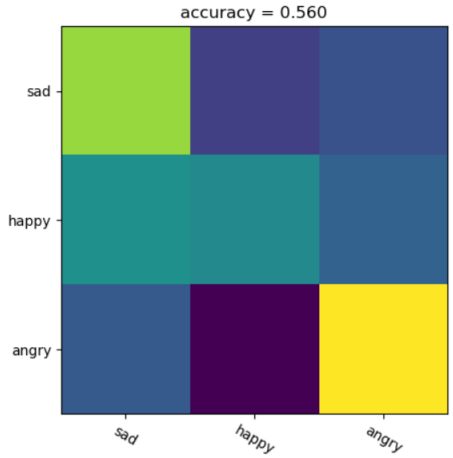



Figure 10. Cat Emotion Detection Result: Our Face Only Method Accuracy and Confusion Matrix

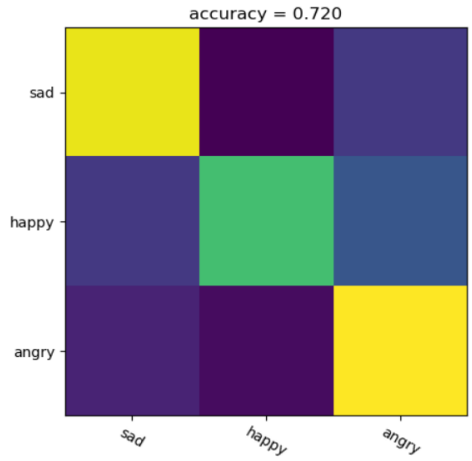



Figure 11. Cat Emotion Detection Result: Our Face Landmark Method Accuracy and Confusion Matrix

In the result above, we can see our second method to recognize cat emotion is the most effective one. Compared with baseline method, our face landmark method can distinguish sad and angry emotion more clearly while the performance of baseline method looks like the random action. The reason may be the HOG descriptor cannot distinguish the edge of cat ears because the fur color of cat is similar to the background, and rotation may destroy features so the SVM cannot classify correctly. Compared with our first method, classifying based on landmark out-performed face only detection because the whole face image may containing some noise and ambiguous information to lower the performance. In the visualization result on test set below, we can clearly see the mouth and eyes are different in different cat emotion.


Figure 12. Our Face Landmark Method Visualization Result

Here we also list some wrong classification result as below. Such mistakes may be caused by low quality of image and ambiguous semantic meaning and patterns.


Figure 13. Our Face Landmark Method Visualization Result: Wrong Classification

## 5. Future Work

We propose the following three aspects as our future work.

1. Facial expression is the most crucial part of detecting cat emotion. However,facial expression can often be ambiguous in pictures. By combining signals from other parts of cat body, such as tail height level, AI can understand cats' emotion more preciously.

2. Due to the lack of cats' emotion data, we manually labeled cat pictures. This may cause bias on our dataset because human cannot fully understand cat emotion just by appearance. Modern techniques and standards, such as AU taxonomy expressed in [6], should be used to correct bias.

3. We discretized cat videos to pictures and detected cat emotion per frame. However, cat emotion should be a continuous state through video. Relation between pictures was neglected. In this case, optical flow would be used in future to capture this relation.

## 6. Conclusion

In this project, our major achievement was a successful emotion recognition on animals, especially cats. We were able to distinguish three types of cats' emotions: angry, happy and sad by using neural network. We used a fine-tuned neural network instead of baseline [HOG + SVM]. By comparing with the result of their methods applied on our dataset, we claimed that our results were more general and accurate. Emotion recognition accuracy was improved from 0.33 to 0.74. Meanwhile, two different training inputs, the entire cats' faces or feature patches extracted from cats' faces were used parallel to train the neural network. It was found that using features as inputs would give more accurate results. Our intuition was that extracted features, such as ears, eyes and mouth, were highly correlated with emotion of cats. Those parts were emphasized while less important parts such as furs were filtered out. At last, we conducted an AI dubbing application, in which a soundless cat miaow video was dubbed with corresponding cat sound based on emotion state of that cat.

## 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] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.

[11] Paul Viola and Michael Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.

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

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

[14] 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.