10-775 Project Proposal
Multi-modal Cartoon Emotion Recognition

Group Members
Liangke Gui, Taiyuan Zhang, Chengliang Lian
liangkeg, taiyuanz, clian

1 Problem Description

Human emotion recognition is an interesting topic in recent studies. Emotions in the movies are detected using machine learning algorithm for further research and commercial use, i.e. the movie classification and recommendation system based on emotions in the movie. Yet seldom researches had been conducted on the emotion recognition of cartoon movies, which differs from the physical emotions in its richness of color, lighting, shape and style. The features could vary a lot from movie to movie. The task is even more challenging when the research goal expands to recognize emotions of personified cartoon characters like cartoon cars.

Our goal is to extract and generalize the feature of emotions in cartoon. Auxillary features from physical word could be applied to the process. Audio feature are also applied in the process of emotion recognition.

2 Data Set

The main data set used for training and evaluation is a cartoon movie data set developed internally in LTI lab. This dataset consists of 6 movies [1]. Each movie is divided into segments of various length, and each segment contains emotion annotation (there're seven emotions in total: happy, sad, surprised, disgust, fear, ...). The task of this project is to predict the label given the feature representation of the video segment.

Considering the lacking of sufficient data in the aforementioned data set, we'll also consider using IEMOCAP[1] as an auxillary training set. More details will be covered at later chapters.

3 Proposed Method

3.1 Feature Design

We plan to investigate the usage of different feature sets so as to improve the final accuracy on our evaluation set. Initially, we're considering the following features[2][4][5]

1. Audio feature, i.e. MFCC
2. Text feature on subtitle
3. Visual features
   (a) color, SIFT, MOSIFT
   (b) Feature learnt by convolutional neural network[6]

[1] Until now, only 4 out of 6 are labeled. We ourselves need to label the rest 2

Taking these as potential initial features, we may also apply feature-fusion techniques, such as multi-stage late fusion.

3.2 Transfer Learning

We also plan to investigate whether it'd be beneficial to introduce auxillary data set and transfer learning technique to deal with lack of sufficient training data. As mentioned above, we might be able to improve the classification accuracy by integrating the knowledge about human emotions extracted from the IEMOCAP dataset. For example, we can use the emotion classifier trained on the IEMOCAP dataset to obtain preliminary audio classifier, apply this classifier on the actual training and test set to obtain a probabilistic distribution over class labels, and introduce this as anther feature.

4 Experiment Design

For experiment design, we plan to divide the dataset into a training set and test set in a movie-wise manner, i.e. all segments from a movie can only either all belong to training set or all belong to test set.

We'll compare the results from choosing different features. Since this is our main focus, we don't plan to spend too much effort on investigating different training models or propose new supervised learning models, therefore we plan to just use Logistic Regression or Kernel SVM as our classifier[3].

5 Plans and Requirement

Some of the features, such as audio and text features, are already extracted within LTI lab, which can be a good start point. We plan our project milestones as follows:

1. Feb 23 - Mar 18: Establish the system pipeline. That is, we can get result from using provided basic features
2. Mar 19 - Apr 15: Develop visual-based features and add into pipeline. Also investigate transfer learning
3. Apr 16 - End: System refinement and parameter tuning, optimization

References

[1] Busso C, Bulut M, Lee C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Language resources and evaluation, 2008, 42(4): 335-359.

[2] Zeng Z, Pantic M, Roisman G I, et al. A survey of affect recognition methods: Audio, visual, and spontaneous expressions[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2009, 31(1): 39-58.

[3] Mower E, Mataric M J, Narayanan S. A framework for automatic human emotion classification using emotion profiles[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2011, 19(5): 1057-1070.

[4] Ververidis D, Kotropoulos C. Emotional speech recognition: Resources, features, and methods[J]. Speech communication, 2006, 48(9): 1162-1181.

[5] El Ayadi M, Kamel M S, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases[J]. Pattern Recognition, 2011, 44(3): 572-587.

[6] Kim Y, Lee H, Provost E M. Deep learning for robust feature generation in audiovisual emotion recognition[C]//Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 3687-3691.