CS 766 Mid-Term Report

Jiangang Chen(jiangang.chen@wisc.edu)

Introduction

This project focuses on the state-of-art Facial Expression Recognition (FER) methods and reimplements the FER algorithms (including Conventional and Deep Learning-based approaches). Two groups will be tested and compared on several datasets (In-the-Lab/In-the-Wild). The project will also discover the performance of two FER methods under infrared conditions.

For more background, please refer to the Project Proposal.

Progress

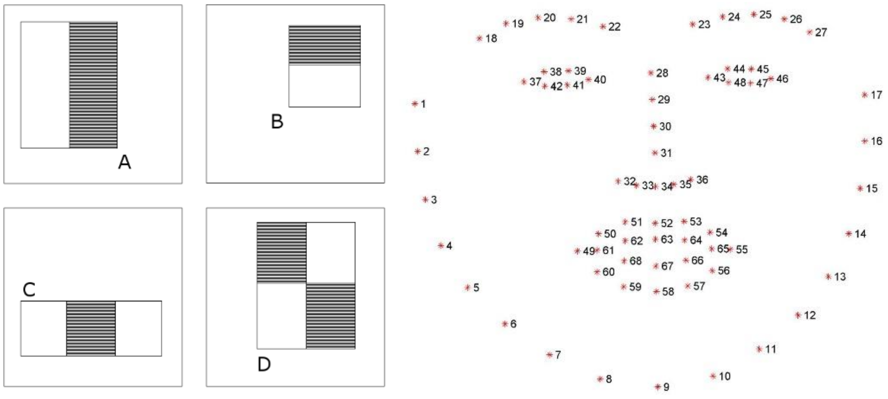
Conventional Method

Revisions

Traditional FER methods include three function blocks. The first function is face detection and facial components extraction. The second function block is using the Histogram of Oriented Gradients (HOG) to encode these components. The last function block is training a Support Vector Machine (SVM) classifier.

As mentioned in the Project Proposal, this project will reimplement the conventional approaches to FER based on the work [1]. However, there isn't any supportive code for this work. So, I will replace the first part of conventional FER with dlib-facial_landmark.

Here, the reason why dlib-facial_landmark can replace Viola-Jones face detector (Haar Cascades) is that once the face region is acquired, both methods are able to "extract" the brows, eyes, nose, and mouth from the face. (Although no "feature extraction" is taking place in dlib)



(A) (B)

Fig. 1. (A) Example rectangle features from Haar Features (b) Facial landmarks shape from dlib "shape_predictor_68_face_landmarks.dat"

Progress

Facial expressions result from muscle movements and these movements could be regarded as a kind of deformation. HOG features are pretty sensitive to object deformations. So the algorithm uses HOG to encode the components. Support Vector Machine (SVM) has been widely used in various pattern recognition tasks. It is believed that SVM can achieve a near-optimum separation among classes. In general, SVM builds a hyperplane to separate the high-dimensional space. The original SVM is a binary classifier. However, we can take the one-against-rest strategy to perform the multi-class classification.

With the newly added dlib face landmarks, HOG, and SVM, we can first compare the performance on the same datasets with different combinations.

The target dataset is Fer2013(in-the-lab dataset), which is a classic dataset, containing 30,000 facial RGB images of different expressions with a size restricted to 48x48, and the main labels of it can be divided into 7 types, 0-6: Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral. The program also tests the 5 types of expression, which same as the 7 types but exclude Disgust and Fear.

As shown in table 1, we can first conclude that the accuracy increases slightly by using the sliding window during HOG. And dlib face landmarks have better performance than the HOG features. Also combining the dlib with HOG (same steps for Haar with HOG), we can improve the performance dramatically.

TABLE I. Accuracy results from different feature combinations

Features	7 Emotions	5 Emotions
HOG Features	29.0%	34.4%
Face Landmarks	39.2%	46.9%
Face Landmarks + HOG	48.2%	55.0%
Face Landmarks + HOG on sliding window	50.5%	59.4%

Future Goal

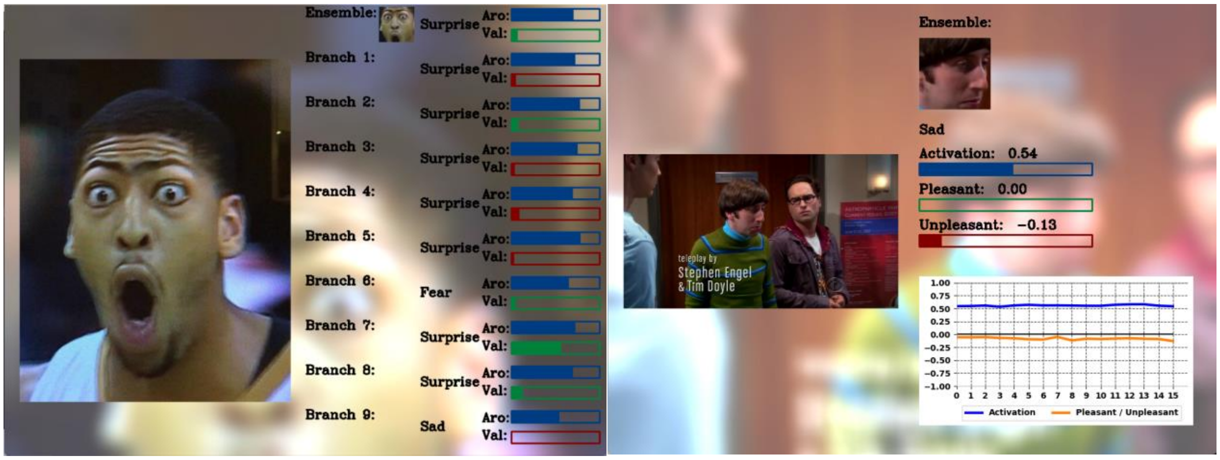
Next, I'll test the convention FER method on another in-the-lab dataset (CK+) and in-the-wild datasets (AffectNet). Then I will compare the results with deep learning-based approaches.

Deep Learning-based

Progress

Similar to conventional methods, the deep learning FER methods detect the faces first. As described before, two methods, dlib-facial_landmark and Haar Cascades, can be used to detect face and extract components. Although, dlib is slower but accurate, whereas Haar cascade is fast but less accurate. The program will use dlib to measure the prediction accuracy, but also provide Haar Cascades to make a fast process in practice applications.

Ensemble methods have proven to be efficient methods for reducing the remaining residual generalization error, which results in robust and accurate methods for real-world applications. In the context of deep learning, however, training an ensemble of the deep network is costly and generates high redundancy. Here the convolutional neural network is an Ensemble with Shared Representations (ESRs), which has high data processing efficiency and scalability to large-scale datasets.



(A)

(B)

Fig. 2. Examples of Deep Learning-based FER methods on image and video. (A) Surprising Expression (B) Sad Expression

Future Goal

Next, I'll test the deep learning-based FER method on the in-the-lab datasets (Fer2013, CK+) and in-the-wild datasets (AffectNet), and compare the results to the conventional FER method.

Timeline Revisions

There aren't too many changes needed to revise the timeline in the project proposal. But during the period from the previous submission, I did find that,

- It's time-consuming to test a variety of datasets especially since each one needs to go through the files rearranging, data loader, and fitting of the program. I'll try to test 2 datasets for each method.
- The website (which will be on GitHub) is under development, it's roughly one week later than I expected. This is because I need to deeply understand paper and algorithms. Also, it's my first time implementing the website on GitHub. It may take some time to do the research on it.
- Testing both methods under Infrared conditions is my interest. If I can finish the benchmark early, I'll try it.

Overall, the timeline could stay the same as before. But I need to spend more time every week finishing the project.

Reference

[1] Chen, Junkai, et al. "Facial expression recognition based on facial components detection and hog features." *International workshops on electrical and computer engineering subfields.* 2014.