FYP17019
2017-2018

Lo Wang Kin
3035186401

Mak Tin Shing
3035187857

Supervisor: Dr K.P. Chan

Facial Expression Recognition using Regional Proposal

Project Plan

Contents

Introduction

Facial Expression has always been an important role in face-to-face communication. Messages sent and received through the use of facial expression are easily understood by human without much effort, but it is not an easy task for computers. Understanding human emotions through facial expression is still a challenge for computer system[1].

In order to facilitate a more natural communication between humans and computer, considerable progress has been made in the field of feature extraction algorithms and classification techniques in the past few years. Much development involves using a techniques called Facial Action Coding System (FACS) [2,3]. FACS is a method that classifies different facial components, such as eyes and lips, into some Action Units (AUs). Combining different AUs, the computer can then recognize the expression. Some other methods involving classifiers like Naive Bayes classifiers and hidden Markov models[4].

The recent success in Convolutional Neural Network (CNN) has draw much attention in the field of Facial Expression Recognition (FER). Different contributors try to apply modifications on CNN such as applying transformation on images at train time, using a decision tree for combining results of different neural networks, and use different pooling layers [5,6,7]. However, most CNN models usually cannot achieve both high recognition rate and high accuracy. Much experiments and researches are needed to examine the success and drawbacks of different techniques.

In order to extend the use of CNN to object detection, the Region-based Convolutional Neural Networks (R-CNN) was developed in 2014 [8]. The system proposes regions using selective search before feeding them to a modified version of AlexNet for classification. The algorithm was further optimized to Fast-R-CNN, Faster-R-CNN and Mask-R-CNN. R-CNN successfully improves the precision on PASCAL VOC 2012 by 30%, compared to CNN [8]. The success of R-CNN could be a key to further improve the performance of FER.

Objective

This goal of this project is to:

1. **Implement the models for FER**
This is the basic part of this project. The main goal of this part is to implement some functional region proposing models for the next part;

2. **Experiment different region proposing techniques**
This is the main objective of this project. This part mainly focus on experimenting different region proposing technique in FER task. The experimental result will be based on the final mean average precision (mAP) and the mAP per training will be compared. The training speed and classifying speed will not be compared as they are heavily dependent on optimization.

Methodology

This project will use Caffe framework for development. Caffe is a popular framework of neural networks, so finding pre-trained model and troubleshooting technical issues would be easier. Also, the modularization of layer enables easy switching of predefined layer, instead of implementing them from scratch. Moreover, the open source project of R-CNN, Fast R-CNN and Faster R-CNN are written in Caffe. So using Caffe minimizes the time needed for translation.
The project involve 3 neural networks. The networks will train on a dataset from Cohn-Kanade database. The dataset contains 593 sequences of images, each of them begins with a neutral expression and proceed to a peak emotions. There are a total of 11424 images with 8 kinds of expressions, namely neutral, angry, contempt, disgust, fear, happy, sadness, and surprise. For each network, training curve and mAP will be recorded for comparison.

The project will start with a traditional CNN using transfer learning and fine tuning. The pretrained model is the BAIR/BVLC CaffeNet Model [9] which is trained on ImageNet dataset managed by the Stanford Vision Lab. This CNN model will be the baseline for comparison.

The second neural network will use a pre-trained Faster R-CNN to extract labelled regions useful for expression classification, like eye area and mouth area. The regions will then be feeded to the baseline CNN for classification. Different strategy of integrating the labelled regions will be tested, such as filling the unbounded region with black.

The final neural network will make use of selective search for regional proposal. The proposed regions will be feeded to the baseline CNN. Again, different strategy of combining the regions will be tested.

Schedule

Sprint 0: Preparation

- Gather requirements — September 2017
- Study the data set — Early October 2017
- Project plan — 1 October 2017

Sprint 1: Baseline model

- Implement the pre-trained CNN model — October - November 2017

Sprint 2: Faster R-CNN model

- Design and implementation of R-CNN model — November - December 2017
- First Presentation — 8-12 January 2018
- Interim Report — 21 January 2018

Sprint 3: Selective search model

- Design and implementation of selective search model — January - February 2018

Sprint 4: Testing and optimization

- Experiment different techniques for combining regions — February - March 2018

Sprint 5: Finalization

- Evaluate test results — Early April 2018
- Final report — 15 April 2018
- Final presentation — 16-20 April 2018
- Project exhibition — 2 May 2018

Reference

[1] N. Sebe et al., Authentic facial expression analysis, Image Vis. Comput. (2006), doi:10.1016/j.imavis.2005.12.021

[2] Y. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2), February 2001

[3] Stewart Bartlett, Marian & Littlewort, Gwen & Frank, Mark & Lainscsek, Claudia & R. Fasel, Ian & Movellan, Javier. (2006). Automatic Recognition of Facial Actions in Spontaneous Expressions. Journal of Multimedia. 1. . 10.4304/jmm.1.6.22-35.

[4] Cohen I., Sebe N., Sun Y., Lew M.S., Huang T.S. (2003) Evaluation of Expression Recognition Techniques. In: Bakker E.M., Lew M.S., Huang T.S., Sebe N., Zhou X.S. (eds) Image and Video Retrieval. CIVR 2003. Lecture Notes in Computer Science, vol 2728. Springer, Berlin, Heidelberg

[5] Kim, BK., Roh, J., Dong, SY. et al. J Multimodal User Interfaces (2016) 10: 173. https://doi.org/10.1007/s12193-015-0209-0

[6] Zhiding Yu and Cha Zhang. 2015. Image based Static Facial Expression Recognition with Multiple Deep Network Learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15). ACM, New York, NY, USA, 435-442. DOI: http://dx.doi.org/10.1145/2818346.2830595

[7] Mollahosseini, A., Chan, D., & Mahoor, M. H. (2015, November 12). Going Deeper in Facial Expression Recognition using Deep Neural Networks. Retrieved September 27, 2017, from https://arxiv.org/abs/1511.04110

[8] Girshick, Ross and Donahue, Jeff and Darrell, Trevor and Malik, Jitendra. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. Computer Vision and Pattern Recognition

[9] Jeff Donahue. (2014). BAIR/BVLC CaffeNet Model. Retrieved from https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet