Facial Expression Recognition and Tracking

Sanmathi Kamath, Advait Koparkar, Pranjali Kokare
Georgia Institute of Technology

Project Summary

The objective of this project is to implement Facial Expression Recognition and Tracking (FERT) in videos. Facial expressions are a powerful cue that convey critical information in almost all human interactions. The ability to understand human facial expressions can equip machines with an incredibly useful window into the mind of humans. In the era of automation, to have a machine that has the ability to extract information from facial expression can be tremendously useful in a wide variety of applications, especially in the ones which involve human-computer-interaction.

The objective of the project is to classify the facial expressions of multiple subjects in a video sequence and to track the changes in the facial expressions of multiple subjects appearing in the video. We propose a machine learning driven pipeline that has the ability to achieve a high-level understanding of facial expressions of multiple subjects in videos. Through this project, we seek to combine a variety of supervised and unsupervised learning methods in an attempt to maximize the accuracy and reliability of the system for the FERT task. In the process, we also aim to develop a deeper understanding of practical machine learning algorithms. We have broken down the task of Facial Expression Recognition and Tracking into 3 well-defined parts:

- **Facial Expression Recognition (FER) from Images**:
  The goal of this task is to train a classifier that will be able to predict facial expressions from images of human faces. We will first address the question of why we can treat FER as a classification task and why the use of machine learning is justified. The classifiers will be trained on a dataset [1] that will classify facial expressions into 1 of 7 categories. Justification for choosing 7 specific classes will also be provided. We will then train various kinds of classifiers including K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Deep Convolutional Neural Networks (DCNN) [2] etc. and analyze the results of classification achieved by each of these methods using various accuracy metrics.
- **Face Detection (FD) in Images**:
  For the system to work on natural images and videos, it will have to identify the parts of natural images which contain faces. The problem of Face Detection is a special case of the more general task of Object Detection. We will use the state-of-the art face detection algorithms [3, 4] which have shown accurate and reliable results for Face Detection.
- **Face Tracking (FT) in Videos**:
  The ability to understand the time-variation of facial expressions of individuals in videos can only be achieved if the system is able to match faces seen in the successive frames of the video. We will implement various unsupervised clustering algorithms like K-Means, Mean Shift, DBSCAN and other variants to achieve this objective. The attempt will be to develop an approach that will yield results that are robust to possible causes of error such as occlusions and change in illumination across the frames of the video sequence.

We intend to combine the 3 steps mentioned above to achieve the task of Facial Expression Recognition and Tracking in videos. The primary motivation for implementing FERT is drawn from its possible applications in today's world. Understanding facial expressions from videos can serve as a much simpler alternative to measuring biological signals (like MRI scans, heart rate etc) to understand the mental state of human beings. Such a system may also find applications in the industry where companies could improve the way they understand how the consumers feel about their products. Robust FERT algorithms have the potential to revolutionize the way we interact with machines. Another motivating factor behind this project is the variety of machine learning techniques that will be combined to achieve the result. Implementing and analyzing the performance of various machine learning techniques will give allow us to develop an intuition and in-depth understanding of various practical machine learning algorithms.

Detailed Project Description

1 Introduction

Facial expressions are a significant part of how we perceive other person's thoughts, which plays a major role in almost all natural settings of human interaction. Today, we depend on machines driven by AI for a wide variety of tasks. Human-Computer Interaction has never been more prevalent in our lives, and its influence continues to grow as we build more robust machine learning applications which automate different tasks. The objective of this project is to propose a machine learning driven pipeline that would be able to achieve a high-level understanding of facial expressions of multiple subjects in videos, and thus develop a Facial Expression Recognition and Tracking (FERT) algorithm. This project involves experimenting with a variety machine learning algorithms - spanning supervised learning and classification to unsupervised learning techniques, from using well-established approaches like K-Nearest Neighbours(KNN) and Support Vector Machines to experimenting with Deep Neural Networks. It is an attempt to implement and combine various approaches to achieve accurate and reliable results while broadening our understanding of various techniques of machine learning. We begin by briefly describing the problem of Facial Expression Recognition and Tracking (FERT) in videos.

2 Problem Definition

We define Facial Expression Recognition and Tracking as a two-part problem - (1) identifying facial expressions from images and (2) tracking the time-variation of facial expressions of multiple subjects in videos. We model the first part as a classification problem where we seek to classify facial expressions into 7 distinct categories. We plan on tackling the second part using unsupervised learning and clustering techniques. A more in-depth and formal problem description is explained in section 4. Before diving into the details of the planned methodology, we provide motivation for undertaking this problem and justifying the use of machine learning techniques in order to solve it.

3 Motivation

In this section, we will address two broad issues: (1) Why the use of Machine Learning techniques is justified for this task and, (2) Why the task of Facial Expression Recognition and Tracking (FERT) is relevant.

3.1 Why use Machine Learning?

A fundamental question that may be asked before trying to classify human facial expressions is - is it reasonable to go about trying to classify and generalize human expressions into a finite set of universal categories? This had been a long-standing debate among scientists and psychologists since the 19-th century. Prominent scientists like Charles Darwin argued that there exist a set of universal facial expressions in humans that are invariant to changes in culture. But it was only in late 20-th century that scientist Paul Ekman presented a quantitative study [5] which concluded the argument in the favor of existence of universal facial expressions. The study concluded by identifying 6 different facial expressions - happiness, sadness, anger, fear, surprise and disgust which the author called "pan-cultural". In this project we work with a dataset [1] which conforms with the result of the above study. The dataset to carry out Facial Expression Recognition(FER), classifies faces into 7 different categories. These include the 6 categories mentioned above and an additional "neutral" facial expression category. Thus, our attempt to generalize facial expressions in humans is justified. The task of FER cannot be broken down into simple quantitative comparisons between hand-crafted features and is a complex task which cannot be solved using naive heuristic approaches. This justifies the use of Machine Learning techniques to tackle the problem of classifying human facial expressions.

3.2 Why Facial Expression Recognition and Tracking (FERT) is relevant?

Understanding human expressions can act as an incredibly useful window into the mind of humans. In today's world which is becoming increasingly automated, a machine having the ability to extract information about facial expression can be tremendously useful in a wide variety of applications which involve human-computer interaction. Traditionally, science has relied on observing biological signals such as heart-rate, temperature dynamics and signals in the human brain. These signals are difficult to observe and thus not feasible to implement in most settings. Using images and videos to perform FERT has the obvious advantage easily capturing the data required to make predictions. Augmenting FERT in videos with other techniques for expressions identification such as speech can be used to build reliable multi-modal systems for human expression recognition and context understanding. In this section we describe a few examples of where the FERT in videos can be implemented:

- Today, a lot of companies use automated calls to customers seeking feedback for products. These automated services use voice and text data acquisition to understand the feedback given by the customers. The process of seeking feedback can be enhanced to gain further and more reliable insight into customers' thought process if FERT is used to analyze the facial expressions when an advertisement introducing a product is shown to customers.
- Due to the ease of storing data and the widespread use of the internet today, we have access to several GigaBytes of video and image data. FERT can be used as a building block for a content-based search engine to make it easier to search for certain categories of videos or images from a database.
- FERT can be used as a tool to refine presentation skills of humans. A system which can analyze the time variation of facial expression can be used to assess the quality of presentations. This can be especially useful in the case of webinars where the audience members are scattered all over the globe. A system that can assess the mood of the audience members can be useful to calibrate the content and quality of presentations.
- Understanding human expressions can also be useful in settings where subjects are being interviewed or interrogated. Understanding the questions which stimulate different facial expressions can help assess the mental state of the person being interviewed. This can find applications in company recruiting and hiring processes and criminal interrogations.

The above mentioned applications require a real-time and robust FERT system for videos. The above applications serve as a strong motivation to apply the concepts of supervised and unsupervised machine learning to study and implement the FERT task.

4 Proposed Approach

This section provides a detailed and formal description of the objective of the project and the three main steps that have been identified to implement Facial Expression Recognition and Tracking (FERT). Given a video sequence, we aim to find a mapping between the faces across the frames of the video. Our final objective is to classify the facial expressions of each subject appearing in all the video frames. This will enable the system to understand how the expressions and emotions of each person changes as the video progresses.

FERT in videos can be broken down into three well-defined tasks. Here is a step-wise explanation of these tasks and brief description of how we intend to make use of machine learning to implement each of them.

- **Facial Expression Recognition (FER) from Images**:
  The problem of Facial Expression Recognition is tackled as a classification problem in this project. Consider an image of a face denoted by f. The classifier will map f to an facial expression $\hat{e} \in \mathscr{E}$, where $\mathscr{E}$ is a set of 7 human expressions that are defined for this task. This classification is carried out using supervised machine learning where we would train a diverse set of classifiers including but not limited to K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Logistic Regression (LR) etc. We will also train a Deep Convolutional Neural Network (DCNN) for this task. We will then compare and

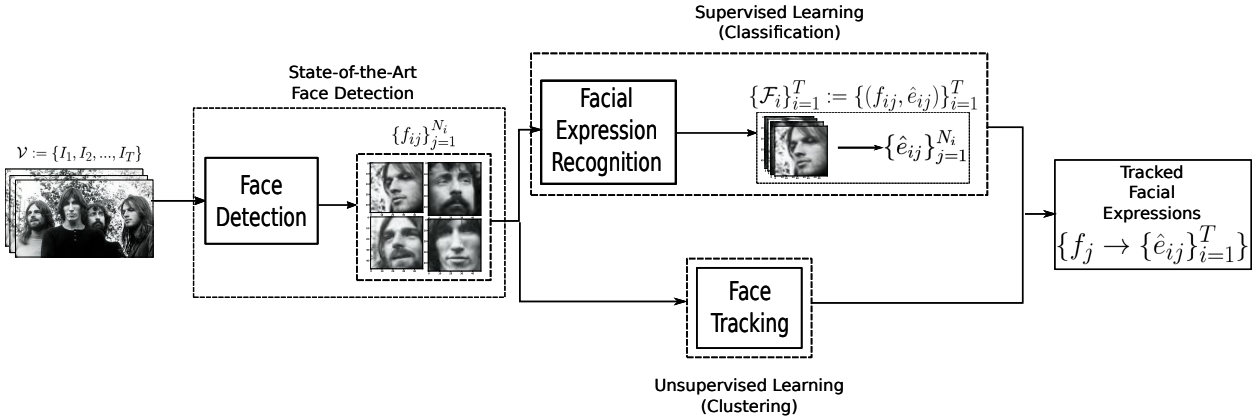

Figure 1: Proposed approach for Facial Expression Recognition and Tracking in Videos

analyze the results of the classifiers based on a set of metrics such as accuracy, F-1 score etc.

- **Face Detection (FD) in Images**:
  Given an image I, the task of the face detector is to estimate a bounding box for each face f_j in the image. Each image may contain a variable number (say N) faces, the face detector's task is to find the co-ordinates of a bounding box (BB_j) for each face. This is a subset of the general problem of Object Detection. Recent advancements in Deep Learning architectures (Faster RCNN, YOLO etc.) have achieved a startling degree of accuracy in the task of object detection. Rather than implementing the Face Detector from scratch, we will use pre-trained architectures and image processing libraries to implement this task.
- **Face Tracking (FT) in Videos**:
  Consider a video sequence $\mathscr{V}$ - a sequence of images $\{I_1, I_2, ..., I_T\}$. Consider the task of face detection and FER are already carried out for these images. So we have a sequence of sets $\{\mathscr{F}_i\}_{i=1}^{T}$, where $\mathscr{F}_i$ is a set of faces mapped to their predicted expressions, that is: $\mathscr{F}_i := \{(f_j, \hat{e}_j) | \forall\, j \in [1, N_i] \text{ and } \hat{e}_j \in \mathscr{E}\}$ where N_i is the number of faces detected in the i-th frame of $\mathscr{V}$. The task of the face tracker is to find a mapping between the faces across the frames of the video. In this project, we will use an unsupervised learning framework to implement FT. We intend to experiment with different types of clustering algorithms (such as K-Means, Mean-Shift, DBSCAN etc.) and their variants in order to achieve robust face tracking. We expect this task to be tricky since the FT algorithm will have to account for various non-ideal situations. For instance, a change in lighting or occlusion can make mapping of faces difficult and error-prone. Since there could be a variable number of subjects in the frames across the video there may be frames which do not have a matching candidate in all the frames of the video. In our attempt, we will try to maximize the reliability and the accuracy of the FT algorithm.

5 List of tasks/collaboration plan

Task	Leader	Deadline	Importance	Potential Challenge
Project Proposal	Pranjali	26th March	Critical	
Background Study	Advait	26th March	Critical	
Data Gathering and Preprocessing	Sanmathi	30th March	Critical	Possible problem : Missing rows, null data Solution: Clean data set with standard techniques
FER (Classification)	Pranjali	6th April	Critical	Problem: Achieving a high accuracy on the data, and also prevent overfitting Solution: Intelligent Model Selection and Cross Validation
Face Detection	Advait	10th April	Critical	Possible problem : Erroneous detection due to occlusion/illumination Solution: Implementing proven, state-of-the-art methods
Face Tracking (Clustering)	Sanmathi	17th April	Critical	Possible problem : Tracking variable number faces in frames in natural images Solution: Carefully choosing the scheme for unsupervised tracking
Combining the pipeline	Pranjali	20th April	Critical	Possible problem : Merging implementations to create a combined code base Solution: Write well-designed, clean implementations from the beginning
Project poster	Sanmathi	26th April	Critical	
Project report	Advait	2nd May	Critical	

Please note each task will have equal contribution from the three of us. We have further subdivided each task to ensure equal contribution.

References

[1] Ian Goodfellow, Dumitru Erhan, Pierre-Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests, 2013.

[2] Chieh-Ming Kuo, Shang-Hong Lai, and Michel Sarkis. A compact deep learning model for robust facial expression recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[3] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *bmvc*, volume 1, page 6, 2015.

[4] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[5] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.