7-2020

Interaction Recognition in a Paired Egocentric Video

Ajeeta Rajkumar Khatri
ak6038@rit.edu

Interaction Recognition in a Paired Egocentric Video

by

Ajeeta Rajkumar Khatri

A Thesis submitted

in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in

Computing Science

Department of Computer Science

B. Thomas Golisano College of Computing and Information Sciences

Rochester Institute of Technology

Rochester, New York

July 2020

MS IN COMPUTER SCIENCE
ROCHESTER INSTITUTE OF TECHNOLOGY
ROCHESTER, NEW YORK

CERTIFICATE OF APPROVAL

MS DEGREE THESIS

The MS degree Thesis of Ajeeta Rajkumar Khatri
has been examined and approved by the
Thesis committee as satisfactory for the
Thesis required for the
MS degree in Computing Science

Zack Butler
Dr. Zack Butler, Professor, Advisor

Ifeoma Nwogu
Dr. Ifeoma Nwogu, Assistant Professor

Xumin Liu
Dr. Xumin Liu, Associate Professor

7/30/2020
Date

Interaction Recognition in a Paired Egocentric Video

by

Ajeeta Rajkumar Khatri

Abstract

Wearable devices and affective computing have gained popularity in the recent times. Egocentric videos recorded using these devices can be used to understand the emotions of the camera wearer and the person interacting with the camera wearer.Emotions affect the facial expression, head movement and various other physiological factors. In order to perform this study we collected dyadic conversations (dialogues between two people) data from two different groups; one where two individuals agree on certain topic and second where two individuals disagree on certain topics. This data was collected using a wearable smart glass for video collection and a smart wristband for physiological data collection. Building this unique dataset was one of the significant contributions of this study. Using this data we extracted various features that include Galvanic Skin Response (GSR) data, facial expressions and 3D motion of a camera within an environment which is termed as Egomotion. We built two different machine learning models to model this data. In the first approach we use an application of Bayesian Hidden Markov model for classifying these individual videos from the paired conversations. In the second approach we use a Random Forest classifier to classify the data based on the Dynamic Time Warping data between the paired videos and individual average data for all the features in individual videos.The study found that in the presence of the limited data used in this work, individual behaviors were slightly more indicative of the type of discussion (85.43% accuracy) than the coupled behaviors (83.33% accuracy).

Acknowledgments

I would first like to express my sincere gratitude towards my advisors Dr. Zack Butler and Dr. Ifeoma Nwogu for providing me with an opportunity to explore my scientific mind and guiding me continuously throughout this journey. I would also like to thank Dr. Xumin Liu for serving on my committee and giving me constructive feedback on my work.

Special thanks to Dr. Hans-Peter Bischof and Cindy Wolfer for their continous support and guidance. I would also like to convey my sincere gratitude to the Computer Science department and all the teachers and advisers who helped me in inculcating a scientific thought process, something that I had always yearned for.

I would like to thank Dr. Joe Geigel for providing the Vuzix glasses for data collection. I would also like to thank my peers Ayush Soni and Jiaxin Wang for their help and support in understanding various Machine Learning concepts and assisting in data collection and all the participants who participated in this study and helped in advancing the scientific research.

Finally, I would like to thank all my friends and family especially my parents Rajkumar Khatri and Sumitra Khatri, Martin Khatri(brother), Simran Khatri(sister) and Samarth Singh for continuously encouraging me to pursue challenging work.

Contents

Chapter 1

Introduction

Wearable devices containing different sensors and recorders have become very popular in the recent years. People have started recording their regular day to day activities and like to analyze the data and use that to bring about changes in the lifestyle. This data ranges from videos, audio and various body parameters like heart rates and galvanic skin response [1].There has been a significant amount of research in evaluating such data using computational techniques such as activity recognition, health monitoring, etc., from the first-person point of view [2] [3] [4] [5] [6].

Many times the facial expressions of the person we are conversing with reveal many more details as compared to the words they use. During our day to day interactions with different people, we observe different facial expressions for different conversations. Our expressions also change based on the expressions of the person we are interacting with. Analyzing the other person's behaviour based on our behaviour from a first person point of view can provide us with various insights about how our behaviour affects the behaviour of the other individuals and vice versa.

Recent studies have been done on Recognizing Micro-Actions and Reactions from Paired Egocentric Videos in the work by Yonetani et. al. [4]. This majorly focuses on recognizing various actions in pairs (such as pointing and sharing a point of attention, gestures and nodding in agreement etc.), also focusing on understanding social interaction dynamics by recognizing micro-actions and reactions. This study has an impact on several first-person vision tasks such as video summarization of social events and assistive systems. This work uses point of view (POV) from both the interacting parties and uses these features to classify the various actions performed by both the individuals in a paired way. Inspired by this work we have explored other modalities apart

from hand gestures and body movement such as facial features which represent emotions, egomotion of camera wearer that represents the head movement and the Galvanic Skin Response that represents the arousal state of an individual in a conversation. These features can be used to classify interactions in a paired conversation.

We also developed a new dataset in order to perform this study. In this work we use the data collected using wearable smart glasses provided by Vuzix. Egocentric videos are recorded using these glasses. These videos are augmented with the GSR data collected using the Empatica wristband which collects physiological signals using sensors. As a part of Egomotion [7] analysis we also analyze the camera motion which represents the head motion of the camera wearer. This motion is then converted into feature representing the head movements in a conversation.

Our works presents a data-driven approach for interaction recognition using a Bayesian Hidden Markov Model(HMM) classification technique and a Random Forest classifier. The Bayesian HMM is trained and tested on individual video data from a paired conversation and classification is done on two classes of interaction i.e. agreement and disagreement using features from three different modalities. The training and testing is done by splitting the data and doing a 5-fold cross validation. The Random Forest classifier is trained and tested on Dynamic Time Warping [8] data calculated for the coupled data and also on average values for each feature for individual data from the paired conversations.

The results from this study can be used to design various systems which can assist users in day-to-day interactions with others. Some of the applications include a real-time system that can analyze when a conversation is changing from an agreeable to a disagreeable one, and alert the user to modify his/her behavior to change the course of the conversation accordingly. Although the constructs investigated here are agreement versus disagreement, many other conversational constructs can be similarly studied. Real-life uses can be found in any face-to-face interactions such as in customer relations, student mentoring, parent-child interactions, employee situations, etc.

1.1 Related Work

Egocentric videos have various applications in the field of Computer Vision. The main challenge in egocentric videos is the invisibility of the camera wearer, the only data available is what the camera wearer sees. Most of the previous work tries to solve this challenge in different ways. Some previous studies

in past have tried to analyze various actions from an egocentric perspective based on what the camera wearer sees. Other studies have focused on egocentric interactions with inanimate objects like various actions performed in the kitchen [3] or a similar setting. These experiments were centered around Egocentric Activity Recognition. The work done by Li et. al. [3] proposes the modeling of gaze estimation and action recognition. They present a model for jointly estimating the gaze and recognizing actions in first person videos. The gaze estimation is done based on the series of actions performed by the camera wearer and his activities are classified based on the surroundings and the objects.

The next level of interaction from a first person perspective is Human-robot interaction where the videos are recorded from robots (first person) point of view [9]. This is different from other forms of human activity classification where the activity is identified by observing the entire scene from the third person point of view. In these types of activity classification the main challenge is the absence of the entire human figure. The activity is recognized by observing what the robot sees, also using certain other factors like hand gestures, head movement, gaze etc. First person activity recognition is essential in various scenarios to provide a robot activity level situation awareness 'during' social and physical interactions. It can help the robot in identifying different situations, be it a friendly human activity such as 'shaking hands with the observer' or a hostile interaction like 'throwing objects to the observer'. In this study the authors propose a way to predict the action before the onset of the action. This can be helpful in preventing various actions when a robot functions in real world. It is also useful in designing a robot which can act differently based on the situation.

To build on top of this interaction, the later studies proposed the understanding and modeling of human-human interaction. This can be a conversation between two individuals or between a small social group. The work by Li et al. [5] models a relationship between camera wearer and the interactor based on the individual action representations of the two persons. They also model a dual relation between two individuals to represent the action and reaction in the interaction. The features are divided into exo and ego features. The surroundings in the egocentric video represent the camera wearer features also termed as the ego features and rest everything in the frame is termed as the exo-feature that represent the actions of the interactor. This research mainly focuses on representing how a relationship between two individuals can be modeled based on each individuals actions. This research was focused on using various hand gestures and the surrounding data in order to model

the interaction. We derive motivation from this study and try to model the interaction using various other features like facial emotions and physiological data.

Similarly, the work proposed by Yonetani et. al. [4] modeled the social interaction between two individuals by recognizing their actions based on the video recorded using head mounted cameras. Their work shows that the first person and second person point of view in a paired egocentric video are complementary. Their main focus is on recognizing micro actions like subtle nods, slight shift in attention or small hand actions. They tried to understand the complexities of social dynamics and model the micro actions and reactions. This work proposes that the understanding of first person actions and reactions in an interaction can be used to understand the social dynamics and can impact various vision tasks such as video summarization of social events.

Our work proposes the idea of modeling human-human interaction in a paired egocentric video using multi modal signals. These multimodal features include facial expressions, GSR (Galvanic Skin Response) data and ego-motion data recorded in an interaction. Similar to recognizing micro actions proposed by Yonetani et. al. [4], we plan to recognize and classify interactions into agreement and disagreement. The actions will be recorded with a wearable camera device worn by both the individuals in facial expressions involve activation of various action units on the face. AFFDEX SDK proposed by Daniel et. al. [10] will be used to find the emotion state displayed by both the camera wearers in an interaction based on their facial expression. The GSR data is collected using the Empatica E4 wrist band. Based on the study by Liu et. al. [11] GSR data can be used as an effective tool in affective computing research field because GSR activity represents different levels of emotional arousal. We also include egomotion data as another feature for recognizing various emotions of from an egocentric point view. The egomotion of the camera represents the motion of the head of the camera wearer. Our research focuses on identifying the importance of this motion in relation to various human emotions. As per our knowledge there is not much work done in the field of mapping egomotion to emotion.

Overall our plan is to use data from multi modal sources such as human facial expression, GSR data and egomotion data to model an interaction between two individuals from an egocentric point of view.

1.2 Objective

Our aim was to understand the dynamics of human behavior in different conversations from an egocentric (first-person) point of view. Our study involved observing the conversation between two people and understanding different factors which influence their behaviour. Our work is intended to help in multiple computer vision related applications where human emotions and behavior affect the outcome of a social interaction. The major part of our study was involved in collecting egocentric videos from different paired conversations and analyzing the behavior based on facial expressions and physiological data. We proposed to model the influence of human emotions from one person to the other. The analysis was performed on temporal data and features extracted were evaluated to obtain an insight into the participants interaction patterns. In order to perform this experiment, we collected paired egocentric video conversation between people with conflicting viewpoints and people with agreeing viewpoints. This was done in order to analyze different emotions which arise during different scenarios. Based on this modeling we classified the videos into the classes of agreement and disagreement. We also planned to come up with a unique dataset of paired egocentric interaction videos to enable this task of human behavior modeling. This dataset was collected based on two categories of interaction that is agreement and disagreement on a certain topic between the two participants.

Chapter 2

Data Collection

2.1 Paired Egocentric Video Dataset

This study depends on having the egocentric conversation data between two participants. Previous datasets were primarily based on egocentric videos where the participants were performing a task. For our experiment, we collected paired videos of dyadic conversations between participants with agreeing and disagreeing viewpoints. In affective Computing, a conversation is classified by the valence in the conversation which is the measure of representing whether it was a positive or negative conversation. Since agreement and disagreement can be linked with this measurement in a conversation we used these two classes of interaction in our dataset. The facial expressions, conversational speech and the GSR data were recorded for each participant. A list of controversial topics in both national and international politics was compiled and the participants were asked to select one topic for which they were strongly in support and another topic on which they strongly disagreed, and would be willing to defend their viewpoints for each topic selected. The topic list was compiled by having discussions with the leaders of various political groups as well as various politically astute international students on campus. Participants were then selected based on the pairings of the topics selected. One individual could, therefore, participate in both a disagreement and an agreement conversation, but not with the same person. Some prospective participants were not selected if pairings could not be made. We had intended to collect data for 15 agreement conversations and 15 disagreement ones, to result in a total of 30 conversations. Due to the covid-19 global pandemic, our college campus was closed and the data collection was suspended in March 2020. To date, experiments involving face-to-face interactions are

Figure 2.1: Interaction between 2 subjects, each wearing the Vuzix glasses to record egocentric video and audio data, along with the Empatica E4 smart wristband, to record galvanic skin response data.

still suspended. Hence, although we were successful in collecting the data for all fifteen disagreement conversations, we could only collect nine of agreement ones, resulting in a total of 24 conversations for analysis. The data collection setup is shown in Figure 2.1.

2.1.1 Participants

A total of eighteen subjects participated in this study. The participants were students from the University who were either affiliated with different political parties or had strong views on certain aspects of international politics. The participants were informed that the data collected would be used for human behavior analysis in various scenarios. Out of 18 participants, 6 participants were female and 12 were male participants.

2.1.2 Apparatus

Each participant was given a pair of Vuzix [12] glasses and Empatica E4 [13] wrist band to wear. Vuzix glasses are wearable smart glasses with an Android based operating system. This smart glass technology is used to capture the video and audio data from an egocentric perspective. The Empatica E4 wrist bands comes with an inbuilt GSR sensor which measures the constantly fluctuating changes in certain electrical properties of the skin. It also collects various other parameters like Blood Volume Pulse (BVP), from which heart rate variability can be derived. We decided to use the GSR values as it is considered

one of the most sensitive and valid markers of emotional arousal [14].

2.1.3 Task

The task is defined as having a 10-minute long conversation with another participant from an agreeing or a disagreeing viewpoint. The participants were briefed about the entire procedure before the study. Consent was obtained from each participant before the study began. The conversation topics were decided beforehand. The discussions took place in a closed room to provide the participants with some level of comfort to freely discuss their viewpoints. There is a small time difference between the two participants initiating the button-click event and this difference is handled by digitally synchronizing the two videos.

We collected 24 pairs of data where one set consisted of 9 agreement conversations and the other, 15 disagreement conversations. The participants in the agreement set were affiliated with the same political party and agreed on the tenets of their party. In the disagreement videos, the participants were affiliated with a different political party and disagreed on certain tenets of the other party. The videos were labeled as agreement and disagreement videos respectively. All the video recordings and other forms of data are stored on a secure drive accessible only by the researchers of this study.

2.1.4 Data Post Processing

During the data collection both the participants were asked to press the start button on both the devices. Since, the time at which this button was pressed varies for each participant, the data is out of sync. Hence the first step in post-processing is to synchronize the video and GSR for each coupled conversation.

For synchronizing the GSR data, we used the initial timestamp data created by Empatica E4 to establish a common start point for both the participants. Additionally, in order to account for the initial arousal state of each individual, we mean-adjusted the GSR data by subtracting the initial GSR reading from all subsequent values for each participant. After this the GSR data is trimmed to represent 10 minutes long data. The data returned by Empatica contains various physiological parameters apart from GSR, since our aim is to use only the GSR data, we extract that data corresponding to both the participants. This gives us synchronized, 10 minutes long GSR reading for both the participants.

For the video data, we used Adobe Premier Pro software to synchronize the two sets of videos in the conversation. The first step is to synchronize the

Figure 2.2: Sample frames from a paired video conversation showing some blurry frames and other frames where the face is fully or partially missing. Each sequence depicts what the other individual is seeing through the Vuzix glasses.

audio data from both the participants. The next step is the synchronization of the videos for each participant which is done using the synced audio files. This step ensures that both the videos have data which is captured for the same time frame during the conversation. After this steps both the videos are trimmed to 10 minute long videos. The next post-processing step is to stabilize the videos to reduce blurring in the frames. Figure 2.2 shows some sample frames as recorded from the Vuzix glasses, from a pair of individuals from the study. It can be seen that some frames are blurred while others are missing some parts of the head/face of the participant. These effects are due to the egomotion of the camera wearer. We observe these patterns across all the participants, the only difference being the intensity and the interval of occurrence. The `sharpen effect` provided by Adobe Premiere Pro is used to reduce the effects of blurriness and the `warp stabilizer effect` is used to stabilize the videos. This effect was provided by a video python module [15].

Chapter 3

Methods

3.1 Data Analysis

The three modalities considered in this study are face-based data obtained from the videos, GSR, egomotion, which represents the head movement of camera wearer in a conversation. Initially each figure was individually examined to determine its relevance to the classification problem. Figure 3.1 top shows the combined egocentric camera motions (or head movements) for two participants in agreement. This shows a pattern of head-nodding which generally represents a *yes* or agreement gesture. Figure 3.1 bottom shows the pattern of the two separate participants in disagreement, we see no specific nodding or head-shaking pattern in this case representing a steady face.

As a second modality, we consider different features extracted from face images. The Affdex SDK [10] provides an interface for processing multiple faces within a video or live stream in real-time. It provides 7 emotion metrics, 20 facial expression metrics and 4 appearance metrics. Also, to represent a common form of expression in digital communication, it provides 13 emojis. The emotion expressions (Anger, Disgust, Fear, Joy, Sadness, Surprise and Contempt) are based on combinations of facial actions. It also provides facial expressions, such as smile, brow furrow, inner brow raise, brow raise, and nose wrinkle, etc based on the facial action units. This facial units emotions coding was built on the EMFACS [16] emotional facial action coding system. The emotion expressions are given a similar score from 0 (absent) to 100 (fully present). As an additional metric, Affdex also provides engagement values as metrics for measuring the emotional experience from a video. This data is obtained by passing each individual video through the Affdex SDK. The SDK returns a csv file containing the values for all the facial emotions and

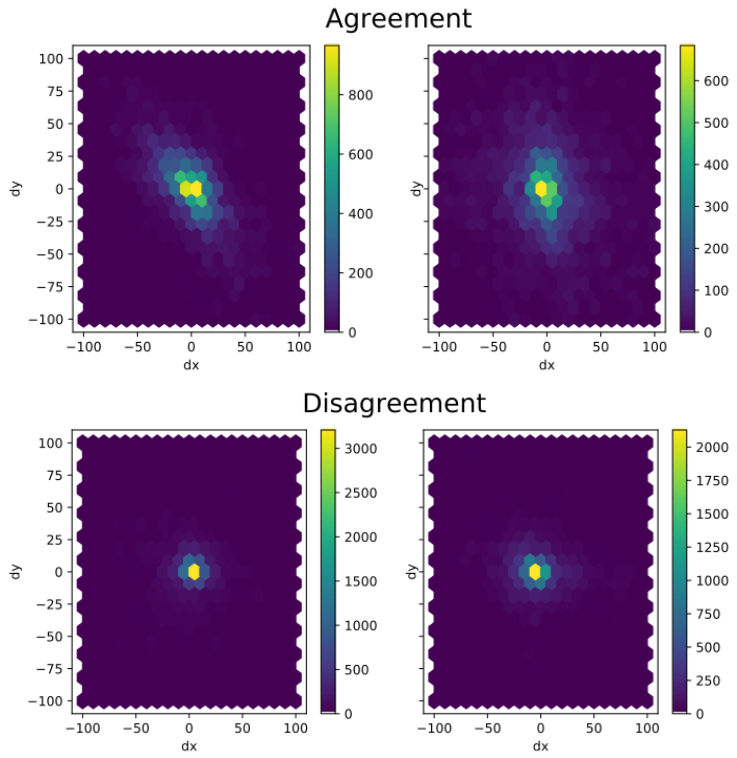
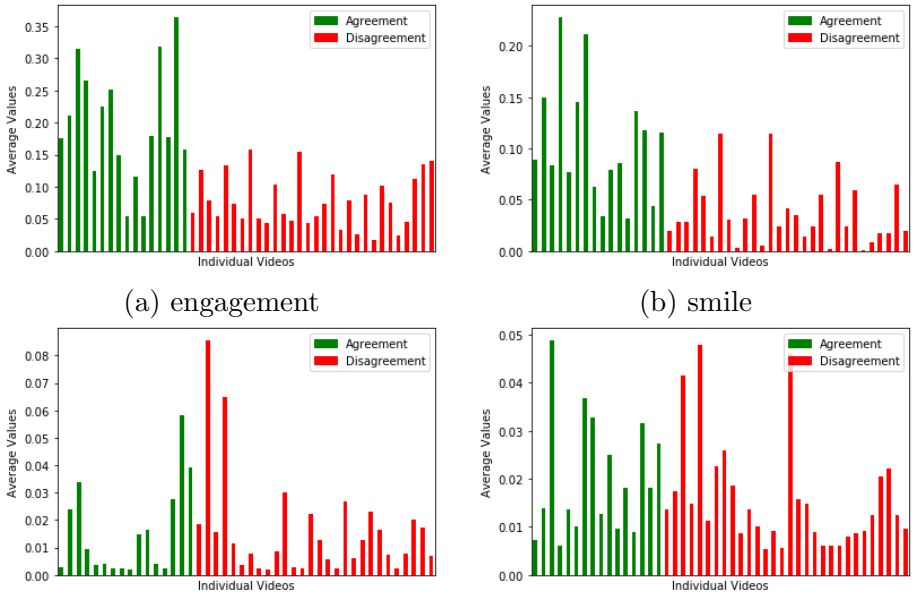


Figure 3.1: Plots showing the *dy* versus *dx* egomotion for 2 pairs of participants. The top images are from the pair of individuals in an agreement conversation and the bottom images are from a disagreement pair.

action units for each frame in the video. Our videos had a frame rate of 24 fps which returned the data corresponding to 14400 frames for each video when passed through the Affdex SDK. Based on the values returned for each video four emotions smile, engagement, contempt and disgust showed higher discrepancies between agreement and disagreement interactions. Figure 3.2 shows the values generated by taking the average of each emotion metric from each participant's video, separating the conversations by their agreement type.

We observe that the distribution of certain facial expressions are more representative of the types of conversations. For example, the overall values for smile and engagement are higher in agreement videos, whereas the values of contempt were significantly higher in the videos where the participants were in disagreement. The distribution of disgust does not appear to be different across the conversation types.



(a) engagement (b) smile

(c) contempt (d) disgust

Figure 3.2: The average values for each participant in a conversation, separated by type

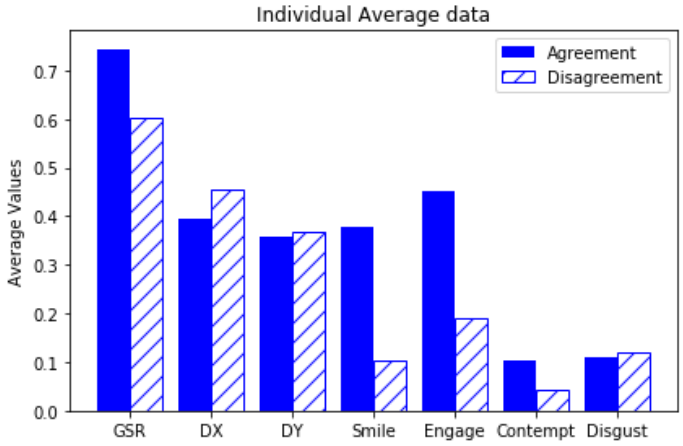


Figure 3.3: Summary of average data for each category

The individual average values for each feature are on a different scale. To compare them on a similar scale we normalize the values for each feature

to have values between [0 1]. To visualize the average value across all the individual videos for each feature, we took the average of the average values of each feature for all the individual videos. As we see in Figure 3.3, shows the normalized average values for all participants in a type of conversation for the features evaluated this work. We observe that the individual traits GSR, smile and engagement have the largest discrepancies between agreement and disagreement.

3.1.1 Coupled Feature Analysis using Dynamic Time Warping

While using the individual traits can provide meaningful data for classification, there may also be value in using in the paired features to determine if they can also play a role in distinguishing between the two types of conversations.

Because all the data collected are time-series data, we are interested in evaluating each pair of sequences when coupled together. One of the common methods of comparing two time series is by performing Euclidean distance matching where the algorithm looks for patterns based on one to one mapping. In these methods, events in the two time series that have similar shapes but different magnitudes, lengths and phases fail to identify correctly. Since our time series data can vary in magnitude and similar patterns can occur anytime throughout the time series, we decided to use the Dynamic Time Warping (DTW) approach which is useful for calculating distance-like similarity measure that allows comparisons of two time-series sequences with varying lengths and speeds. This method is widely known to the speech processing community and we use it for analyzing the similar patterns across time series data for different features.

Timestamp	Participant 1	Participant 2	Common Subset Indices
1	Frame 1	Frame 1 →	1
2	Nan	Frame 2 →	-
3	Frame 3	Frame 3 →	3
4	Frame 4	Nan →	-
5	Frame 5	Frame 5 →	5
6	Frame 6	Frame 6 →	6
.	.	.	.
.	.	.	.
.	.	.	.
N	Frame N	Frame N →	N

Figure 3.4: Common frames are selected from two videos in a pair

Of the three modalities evaluated, we find that the data returned from

Affdex has missing values due to partial or no face detection in a frame. To ensure consistency of data for both the participants, we consider only the data that is present for both the participants at the same timestamp. Since there is no missing data in the case of the other two modalities (GSR and egomotion) we consider the entire data for each pair of conversations. We perform a baseline DTW alignment using the fast_dtw() method provided by the DTAIdistance library [17] and calculate the distance $D(x, y)$ for each feature pair within the two types of conversations. Figure 3.4 shows how common frames are selected from the pairs of videos.

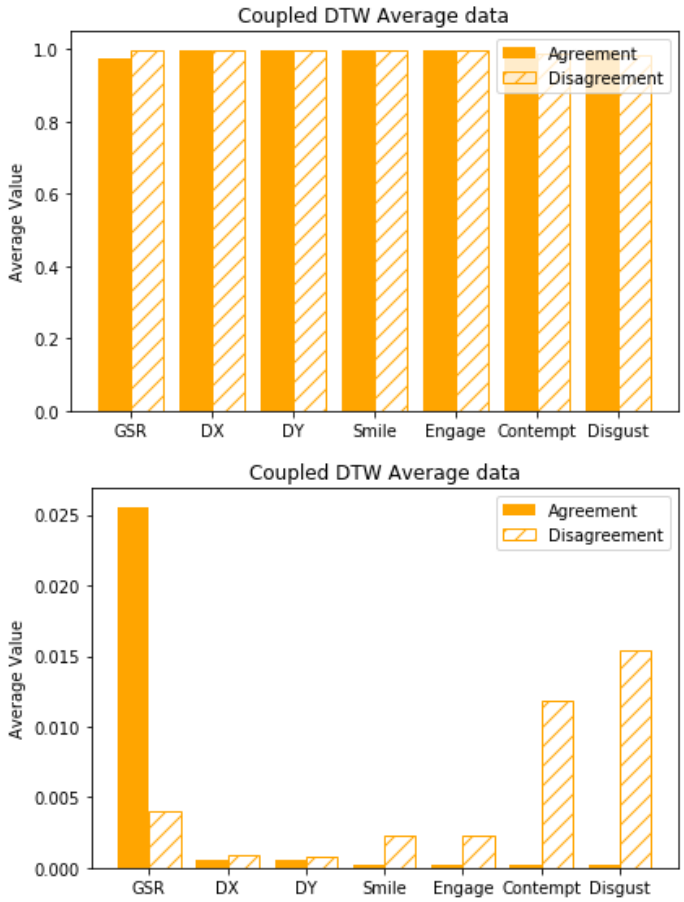


Figure 3.5: Average Similarity(top) and Dissimilarity(bottom) value computed from the pairs of data, for agreement and disagreement conversations.

Since we have different lengths for each pair of conversations, we normalize

the alignment distance by finding the percentage similarity, as given by the formula in Equation 3.1.

$$S(x,y) = \frac{M(x) - D(x,y)}{M(x)} \tag{3.1}$$

Where, $D(x,y)$ is the distance between two time series x and y, S is the normalized similarity measure between x and y, and M is the maximum value that $D(x,y)$ can obtain. In the case of dynamic time warping, given a template x, one can compute the maximum possible value of $D(x,y)$. This will depend on the template, so M becomes $M(x)$. To calculate $M(x)$, multiply the length of the template (the number of samples in the time series), times the maximum value of each sample. If each sample is in the range [0,1], then the maximum value of each sample is 1, so $M(x)$ becomes just the length of x. This gives a simple upper bound on $D(x,y)$. It is easier to verify this data as the resultant $S(x,y)$ is always between the range [0,1].

After calculating the normalized DTW distance between all the pairs in the dataset, the average results obtained for all the features are presented in Figure 3.5. These represent the similarity (top) and dissimilarity (bottom) percentage between the two temporal sequences of the two individuals in a discussion. It shows how some features such as GSR, disgust and contempt have significant differences between the two conversation types. We will use this data to classify the paired data and discuss the results in the ensuing section.

3.2 Bayesian Hidden Markov Model

Hidden Markov Model (HMM) is a statistical model in which the system is assumed to be a Markov process with hidden states and observations. While the states are hidden, the only thing directly visible are the observations. In our case the hidden states represent the emotional state of an individual which leads to emitting the observations which are visible in different forms like facial expression, camera movement termed as egomotion and the GSR data. There are multiple implementations of Hidden Markov models, in most of them the inference is traditionally done using the Estimation Maximization(EM) algorithm. Since the EM algorithm is best described as a method for point estimate, we used Bayesian HMM implemented through Markov Chain Monte Carlo(MCMC) sampling instead of EM algorithm to capture the uncertainty in our parameter estimates. We will now define all the different elements of a HMM using similar notations as used by Dorj [18] and Rabiner [19] in the

table below.

$S_1, S_2...S_K$	a set of K hidden states at any time
$t = 1, 2, ...T$	the time instance
$Q = q_1, q_2...q_T$	a set of hidden states with respect to time. q_t being the state at time t
$A = a_{11}...a_{ij}...a_{NN}$	transition probability represented by $a_{ij} = P(q_t = S_j \| q_{t-1} = S_i)$ and $a_{ij} \geq 0$ and $1 \leq i; j \leq K$.
O=$\{o_1, o_2, ...o_T\}$	a sequence of T observations, each one drawn from a vocabulary V = $\{v_1, v_2....v_m\}$
$B = \{b_j(v_m)\}$	a sequence of observation likelihoods, also called emission probabilities, each expressing the probability of an observation o_t being generated from a state i
$\pi = \pi_1, \pi_2, ...\pi_K$	an initial probability distribution over states. π_i is the probability that the Markov chain will start in state i given by $\pi_i = P(q_1 = S_i)$, $1 \leq i \leq K$.

In compact form the complete parameter set of the HMM can be represented as shown in Eq. 3.2

$$\theta = \{A, B, \pi\} \tag{3.2}$$

Hidden Markov models fall in a subclass of Bayesian networks known as dynamic Bayesian networks, which are Bayesian networks for modeling time series data. In time series modeling, the assumption that an event can cause another event in future, but not vice-versa, simplifies the design of the Bayesian network.. A first-order hidden Markov model instantiates two simplifying assumptions.First, as with a first-order Markov chain, the probability of a particular state depends only on the previous state:

$$MarkovAssumption : P(q_i|q_1...q_{i-1}) = P(q_i|q_{i-1}) \tag{3.3}$$

Second, the probability of an output observation o_i depends only on the state that produced the observation q_i and not on any other states or any other observations:

$$OutputIndependence : P(o_i|q_1...q_i, q_T, o_1..., o_i, ...o_T) = P(o_i|q_i) \tag{3.4}$$

Given this one-to-one mapping and the Markov assumptions expressed in Eq. 3.2, for a particular hidden state sequence Q = $q_0, q_1, ...q_T$ and an observation sequence O = $o_1, o_2, ...o_T$, the likelihood of the observation sequence is

$$P(O|Q) = \Pi_{i=1}^{T} P(o_i|q_i) \tag{3.5}$$

And the joint probability of the being in state Q and generating a particular sequence O is given by

$$P(O,Q) = P(O|Q) \times P(Q) = \Pi_{i=1}^{T} P(o_i|q_i) \times \Pi_{i=1}^{T} P(q_i|q_{i-1}) \tag{3.6}$$

A Bayesian statistical approach to learning and inference starts with some sort of prior knowledge about the model. This initial knowledge is represented in the form of a prior probability distribution over model parameters, and is updated using the observed data to obtain a posterior probability distribution over models and parameters. If we assume a prior distribution over model $P(D)$ and a prior distribution over parameters of each model structure $P(\theta|D)$, a dataset of observations O is used to form a posterior distribution over the models using Bayes rule

$$P(D|O) = \frac{\int P(O|\theta, D)P(\theta|D)d\theta P(D)}{P(O)} \tag{3.7}$$

which averages over the uncertainty in the parameters. For a given model structure, we can compute the posterior distribution over the parameters as:

$$P(\theta|D, O) = \frac{P(O|\theta, D)P(\theta|D)}{P(O|D)} \tag{3.8}$$

In the standard HMM, the state space of the hidden variables is discrete, while the observations themselves can either be discrete (typically generated from a categorical distribution) or continuous (typically from a Gaussian distribution). For an HMM with N hidden states an observation sequence of T observations, there are N^T possible hidden sequences. For real tasks, where N and T are both large, N^T is a very large number, so we cannot compute the total observation likelihood by computing a separate observation likelihood for each hidden state sequence and then summing them.

Instead of using such an extremely exponential algorithm, we use an efficient O(N^2T) algorithm called the forward algorithm described in [20]. The forward algorithm is a kind of dynamic programming algorithm, that is, an algorithm that uses a table to store intermediate values as it builds up the probability of the observation sequence. The forward algorithm computes the observation probability by summing over the probabilities of all possible hidden state paths that could generate the observation sequence, but it does so efficiently by implicitly folding each of these paths into a single forward α value.

Each cell of the forward algorithm $\alpha_t(\text{j})$ represents the probability of being in state j after seeing the first t observations. The value of each cell $\alpha_t(\text{j})$ is computed by summing over the probabilities of every path that could lead us to this cell. Formally, each cell expresses the following probability:

$$\alpha_t(j) = P(o_1, o_2, ...o_t, q_t = j|\theta) \tag{3.9}$$

Here, q_t = j means "the t^{th} state in the sequence of states is state j". We compute this probability $\alpha_t(\text{j})$ by summing over the extensions of all the paths that lead to the current cell. For a given state q_j at time t, the value $\alpha_t(\text{j})$ is computed as

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(o_t) \tag{3.10}$$

With the forward we calculate the maximum log probability and also train the HMM parameters. Now given a new observation sequence we use the forward algorithm and compute the log-likelihood of the observed data, given the model. This method performs computations in log-space to avoid underflow issues and computes and returns the full forward matrix, and the final sum-of-all-paths probabilities.

In our case, the states are hidden and the observations are visible in the form of different features. When using HMM as a classifier, we train one HMM per class for each feature. Considering the states follow a Categorical distribution, we model the transition probabilities as Dirichlet distribution which the conjugate prior to Categorical. Given the continuous nature of the observation we use a Gaussian likelihood and Inverse Gamma prior to model it. After training each HMM, maximum log probability is calculated for each test data point using the forward algorithm discussed above. The HMM which returns a higher log probability is assigned as the predicted class.

3.3 Random Forest

Random Forest is an ensemble-based learning algorithm which is comprised of n collections of de-correlated decision trees. Random forest uses multiple decision trees to average or compute majority votes in the terminal leaf nodes when making a prediction. Built off the idea of decision trees, random forest models have resulted in significant improvements in prediction accuracy as compared to a single tree by growing 'n' number of trees; each tree in the training set is sampled randomly without replacement. Decision trees consist simply of a tree-like structure where the top node is considered the root of the

tree that is recursively split at a series of decision nodes from the root until the terminal node or decision node is reached. A Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. A random forest with dataset X and prediction Y is shown in Figure 3.9 as below.

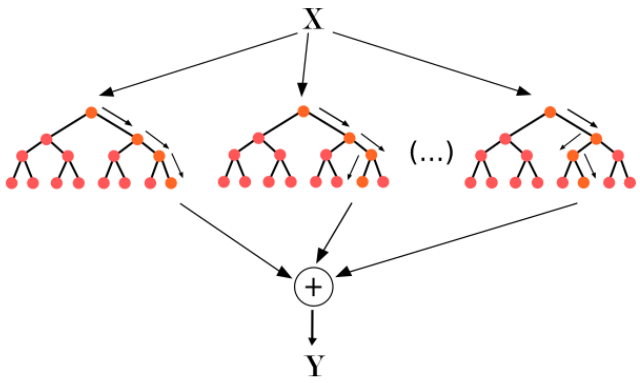


Figure 3.6: Random Forest [21]

Consider a training set with training data $T_1, T_2, T_3...T_n$ corresponding labels as $C_1, C_2, C_3...C_n$, based on the algorithm, random forest may create n decision trees randomly taking input of subset from the training data. Sometimes the data in the subsets may overlap. After training the decision trees and based on the majority of votes from each of the decision trees, it gives the final prediction for the test data. The Random Forest Classifier provided by the scikit-learn [22] library provides multiple options to control different hyper parameters of the model. The n_estimators property provides the functionality to control the number of trees in the forest. And as compared to the traditional implementation the scikit-learn implementation combines classifiers by averaging their probabilistic prediction, instead of letting each classifier vote for a single class. Apart from the prediction it is also important to know which features lead to the prediction. The feature importance ranking can help us identify the important features and also identify the not so important features which can be either removed or processed further to add value to the model.The scikit-learn implementation of Random Forest Classifier provide impurity based feature importance.The higher the value, the more important the feature. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the *Gini*

Impurity. [22]. Gini Impurity measure or Mean Decrease in Impurity (MDI) calculates each feature importance as the sum over the number of splits that include the feature, proportionally to the number of samples it splits. [21].

Chapter 4

Experiments and Results

4.1 Bayesian Hidden Markov Model for Interaction Recognition

Hidden Markov Model (HMM) is a statistical model in which the system is assumed to be a Markov process with hidden states and observations. While the states are hidden, the only thing directly visible are the observations. In our case the observations are visible in different forms like facial expression, egomotion and GSR data and the states represent different emotional states of an individual which are not directly visible. In this study, we chose to investigate each feature independently using Bayesian HMM. For each feature, we created one HMM to represent each class using the data from several individuals. These HMMs were then used for prediction on the test data by computing the probability of the observations in the test sample for each HMM, and predicting class labels based on which HMM returned the higher log probability value. All seven features, one from GSR data, two from Egomotion data (deltaX and deltaY) and four features from facial emotions(smile, engagement, disgust and contempt) were tested in this fashion.

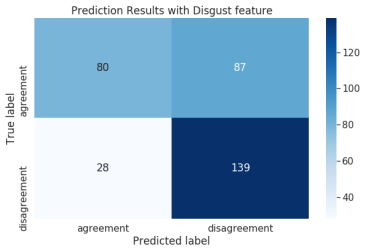
To ensure that no data from the same person was included in the train and test split, we made the split based on individuals rather than data points. For the Bayesian HMM, we considered an equal number of agreement and disagreement videos for training and testing. We took 18 samples from agreement and 18 from disagreement. To perform the split, we did 5-fold cross-validation by splitting the data into 14 conversations for training and 4 conversations for testing.

Each 10 min long individual video was split into 36 video snippets of 16 seconds each, to obtain more data points. Previous research in nonverbal be-

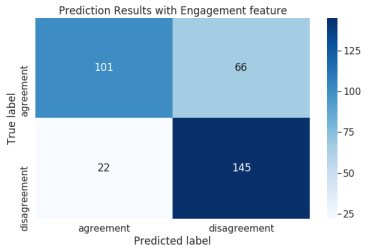
haviour and communication has shown that it takes only 7 seconds to form an opinion about the other person in the conversation [23]. But since our data consists of a paired conversation, we decided to split to videos in 16 seconds each to ensure we could capture enough data from both the participants in the conversations. Each of the video snippets was considered as an individual data point to the model. As Enders [24] stated that a missing data of 15% to 20% was common in psychological studies, we discarded any video snippet which had more than 20% missing data. For the remaining snippets, the data for any missing frames was interpolated using the Impyute library via the Last Observation Carried Forward method. Due to the removal of data points that had less than 80% data, we had some imbalance in training and testing set. The HMM model was trained and tested using 5-fold cross-validation. Figure 4.1 shows four confusion matrices obtained from running Bayesian HMM classifier on the facial features.

The Bayesian HMM was implemented with PyMC3, a python library focused on solving various problems related to Bayesian analysis. For determining the number of hidden states, we used the maximum log-likelihood criteria that best fits the model. For each of the given modalities we ran multiple iterations with different number of states (ranging from 2 - 5) and chose the number of states that maximize the log-likelihood for each modality. Based on the log-likelihood values, we use 3 hidden states for HMMs corresponding to the facial features and 2 hidden states for other two modalities (GSR and egomotion). The transition probabilities between states were specified as Dirichlet distributions. Since our observations are real-valued numbers, we use a Gaussian likelihood and an Inverse Gamma prior to model them. We use the forward algorithm to compute the maximum likelihood of the given observation sequence from training data and train the parameters of the HMM. MCMC sampling methods are used for drawing a series of correlated samples that will converge in distribution to the target distribution. PyMC3 provides No-U-Turn Sampler (NUTS) which is an extension of Hamiltonian Monte Carlo (HMC) to provides an approach for adaptively finding a good number of steps.

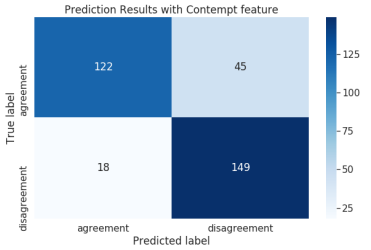
Using this model, we obtain the highest accuracy at about 79% for classifying an interaction into agreement and disagreement by using the contempt feature. The complete list of accuracies obtained is given in Table 4.1



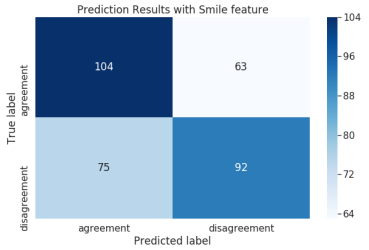
(a) Disgust



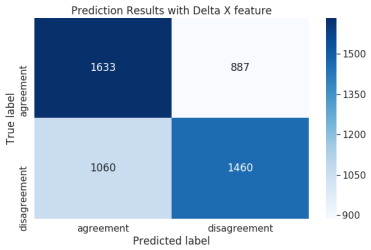
(b) Engagement



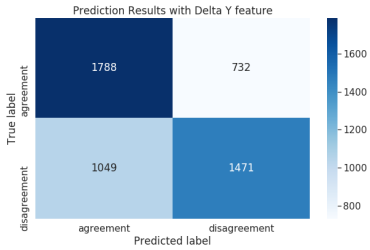
(c) Contempt



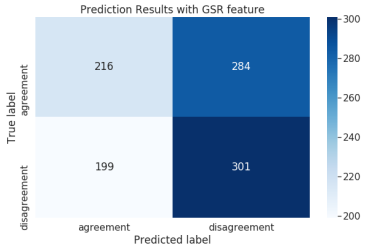
(d) Smile



(e) DX egomotion



(f) DY egomotion



(e) GSR

Figure 4.1: Confusion matrices from running the Bayesian HMM on the different features features.

Table 4.1: Bayesian HMM 5-fold cross validation Average Accuracy

Feature	% Accuracy	Std Dev
Contempt	79.10	13.14
Engagement	75.05	5.71
Disgust	66.80	12.13
Delta Y	64.65	8.31
Delta X	61.36	3.45
Smile	59.77	3.43
GSR	51.7	4.60

4.2 Random Forest Classifier for interaction recognition

Next, we present a different, more discriminative approach to classify the conversations using both individualistic and paired features explained in Sections 3.1 and 3.1.1. The number of data points for both the classifier are very small, 24 in case of coupled video and 48 in case of individual videos. For the individual classifier, each data point represents the combination of average values for all 7 seven features for each individual video. Whereas for the coupled classifier, each data point is the collection of percentage similarity value for 7 features for each coupled video.

In random forests, each tree in the ensemble is built from a random sample drawn with replacement from the training set. A subset of features is considered for splitting each node, which is set to the default value for our model given by `sqrt(n_features)`. The best splitting node is selected by using the Gini Impurity measure. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes the model's prediction. In order to validate the model for such small dataset we use the leave-one-out cross validation which ensures the model is validated on the entire dataset. For each set of features, using the leave-one-out validation method, all the features for all except one participant are normalized and presented to a random forest (RF) classifier to learn how to discriminate between agreement and disagreement interactions. The left-out sample is then tested against the trained RF classifier to yield a result of 1 or 0. This is repeated until all samples have been tested. We developed the RF classifiers using the scikit-learn library [22].

`n_estimators` is one important hyperparameter for training the random forest which defines the number of trees in the forest. For training our model,

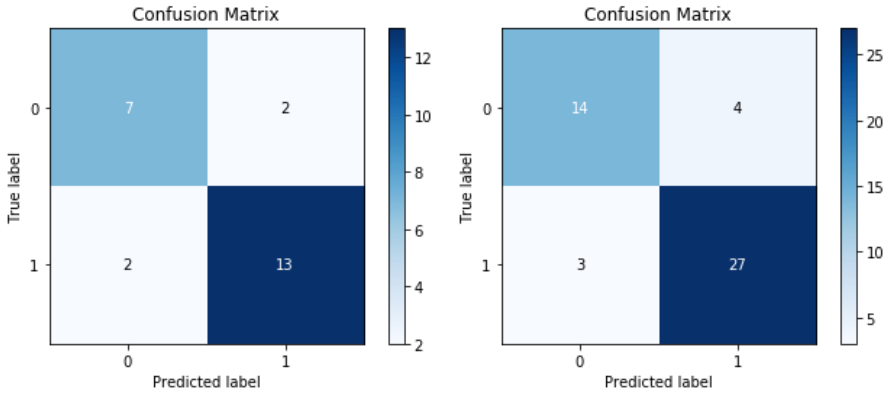


Figure 4.2: The left image shows the confusion matrix for the individualist features (accuracy of 85.43%)and the right matrix is for the paired ones (accuracy of 83.33%).

where the dataset is very small, we `n_estimators` = 10. After performing leave-one-out cross-validation on the data, we obtain the results as 83.33% accuracy for coupled data and 85.43% for individual data. The confusion matrices obtained from testing with the individualistic and paired features are given in Figure 4.2.

Even though we get comparable accuracy with the individual and coupled data for out dataset, the differentiating factor is importance of features for each category. For feature importance ranking, we use the Gini Impurity measure.

Figure 4.3 shows the importance scores for each feature. It is seen that in a coupled video GSR data is the biggest contributor which suggests that the participants in agreement show similar GSR patterns but different from those in disagreement for coupled interactions. This is followed by disgust and engagement pairings. This further suggests that the way the arousal levels of individuals in a disagreement discussion is different from those of the individuals in an agreement interaction.

From the results on individualist features, we observe the facial expression-based features such as smile and engagement are the most prominent determinants. Surprisingly, smile has little influence when the features are individualistic. As shown in both the RF models, we get accuracies between 83-86% for individual and coupled data.

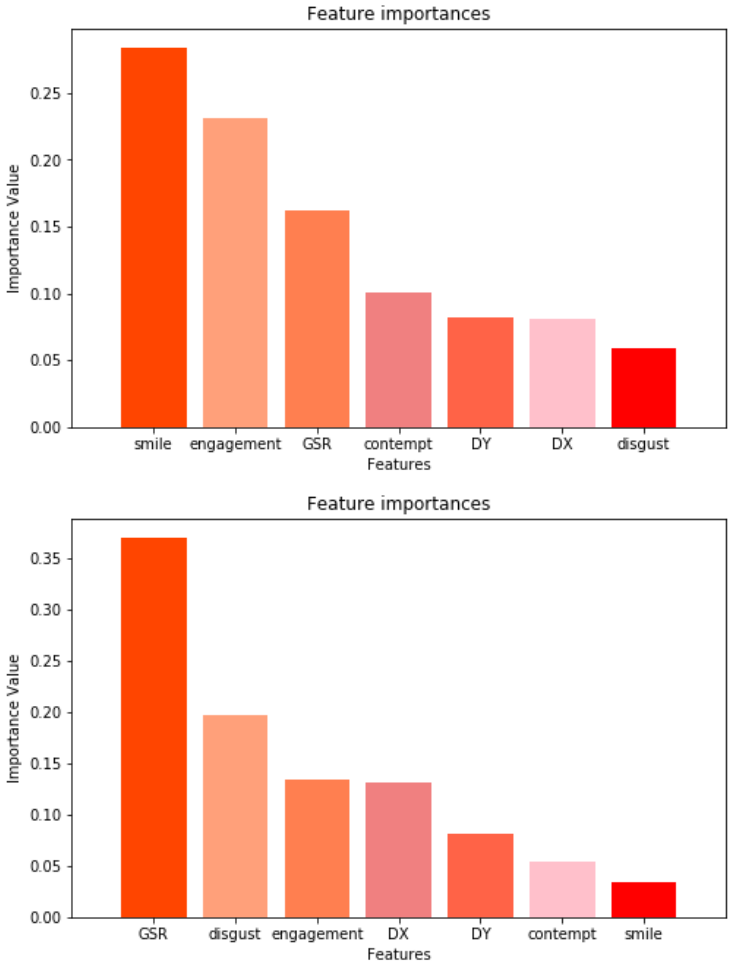


Figure 4.3: Feature importance for individualistic features (top) and coupled features (bottom), for interaction recognition

Chapter 5

Conclusion and Future work

We have presented two learning techniques to model the individual and paired data recorded when two individuals are in agreement or disagreement. A first approach using a Bayesian HMM model to classify the individual videos from a conversation was implemented but this did not yield very strong results. With the current implementation and the high dimensionality of the features, the HMM model did not converge for multivariate features. It would be interesting to observe the behavior of the HMM using multivariate features using an alternate approach. The second approach was to train an RF classifier on both coupled and individual data, where the performance accuracy is 83% and 85% respectively. In the coupled instances, we extracted the alignment scores between sequences and used these as the coupled features. As we analyze the feature importance we observe that features such as GSR, disgust and engagement contribute primarily to the classification of coupled conversations.

It appears that GSR and engagement are the most common and informative feature, given that it plays an important role coupled and individual classifications respectively. In future, this work can be applied to a real time system by considering the features obtained by the camera wearer corresponding to his/her conversant. Affdex provides the functionality to analyze videos in real time and return the facial emotions. Based on our approach for individual video classification the interaction can be recognized from first person's point of view. There can be many such applications which can be designed based on the results and analysis obtained from this study and used in various human-human interaction. The current dataset is very small due to the interruption in data collection during the COVID-19 pandemic. In the future, we would like to collect more varied data from different individuals conversing on a variety of topics and similarly analyze them to draw stronger conclusions.

Bibliography

[1] Galvanic Skin Response. https://www.brainsigns.com/en/science/s2/technologies/gsr.

[2] Suriya Singh, Chetan Arora, and C. V. Jawahar. First person action recognition using deep learned descriptors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[4] R. Yonetani, K. M. Kitani, and Y. Sato. Recognizing micro-actions and reactions from paired egocentric videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2629–2638, 2016.

[5] Haoxin Li, Yijun Cai, and Wei-Shi Zheng. Deep dual relation modeling for egocentric interaction recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[6] Minlong Lu, Danping Liao, and Ze-Nian Li. Learning spatiotemporal attention for egocentric action recognition. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[7] Wikipedia contributors. Visual odometry — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Visual_odometry&oldid=950815719, 2020. [Online; accessed 6-July-2020].

[8] Wikipedia contributors. Dynamic time warping — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Dynamic_time_warping&oldid=965278776, 2020. [Online; accessed 6-July-2020].

[9] M. S. Ryoo, T. J. Fuchs, L. Xia, J. K. Aggarwal, and L. Matthies. Robot-centric activity prediction from first-person videos: What will they do to me? In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 295–302, 2015.

[10] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. Affdex sdk: A cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, page 3723–3726, New York, NY, USA, 2016. Association for Computing Machinery.

[11] M. Liu, D. Fan, X. Zhang, and X. Gong. Human emotion recognition based on galvanic skin response signal feature selection and svm. In *2016 International Conference on Smart City and Systems Engineering (ICSCSE)*, pages 157–160, 2016.

[12] Vuzix. Vuzix. `https://www.vuzix.com/`, 1997.

[13] Empatica. Empatica. `https://e4.empatica.com/e4-wristband`.

[14] Kalliopi Kyriakou, Bernd Resch, Günther Sagl, Andreas Petutschnig, Christian Werner, David Niederseer, Michael Liedlgruber, Frank Wilhelm, Tess Osborne, and Jessica Pykett. Detecting moments of stress from measurements of wearable physiological sensors. *Sensors*, 19(17):3805, Sep 2019.

[15] Python Video Stabilization with OpenCV. Python video stabilization with opencv. `https://pypi.org/project/vidstab/`.

[16] W. Friesen and P. M. Ekman. Emfacs-7:emotional facial action coding system. In *Unpublished manuscript, University of California at San Francisco 2: 36.*, 1983.

[17] wannesm, khendrickx, wusai2333, Toon Van Craenendonck, and Eric Ma. wannesm/dtaidistance v1.2.2, July 2019.

[18] E. Dorj, C. Chen, and M. Pecht. A bayesian hidden markov model-based approach for anomaly detection in electronic systems. In *2013 IEEE Aerospace Conference*, pages 1–10, March 2013.

[19] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[20] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, USA, 1st edition, 2000.

[21] Ramraj Chandradevan. Random forest learning-essential understanding. `https://towardsdatascience.com/random-forest-learning-essential-understanding-1ca856a963cb`, 2017.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[23] Michela Rimondini, Maria Angela Mazzi, Isolde Martina Busch, and Jozien Bensing. You only have one chance for a first impression! impact of patients' first impression on the global quality assessment of doctors' communication approach. *Health Communication*, 34(12):1413–1422, 2019. PMID: 29995443.

[24] Enders C. K. Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. In *Psychological methods, 8(3), 322–337.*, 2003.