Instituto Tecnológico de Costa Rica

Escuela de Ingeniería Electrónica

Automatic detection of human attraction using non-intrusive multimodal signals and machine learning

Documento de tesis sometido a consideración para optar por el título de
Máster en Ingeniería Electrónica con el grado académico de Magister Scientiae

Ronald Gerardo Caravaca Mora

Cartago Agosto, 2022

Declaro que el presente documento de tesis ha sido realizado enteramente por mi persona, utilizando y aplicando literatura referente al tema e introduciendo conocimientos y resultados experimentales propios.

En los casos en que he utilizado bibliografía he procedido a indicar las fuentes mediante las respectivas citas bibliográficas. En consecuencia, asumo la responsabilidad total por el trabajo de tesis realizado y por el contenido del presente documento.

Ronald Gerardo Caravaca Mora

Cartago, 26 de agosto de 2022

Céd: 5-0382-0323

Instituto Tecnológico de Costa Rica
Escuela de Ingeniería Electrónica
Maestría Académica en Electrónica
Trabajo Final de Graduación
Tribunal Evaluador
Acta de Aprobación de Tesis

Defensa del Trabajo Final de Graduación
Requisito para optar por el título de Máster en Ingeniería Electrónica
Grado Académico de Magister Scientiae

El Tribunal Evaluador aprueba la defensa del Trabajo Final de Graduación denominado **"Automatic detection of human attraction using non-intrusive multimodal signals and machine learning"**, realizado por **Ronald Gerardo Caravaca Mora** Carné: **2019390165**, y hace constar que cumple con las normas establecidas por la Unidad Interna de Posgrados de la Escuela de Ingeniería Electrónica del Instituto Tecnológico de Costa Rica.

Miembros del Tribunal

JOSE PABLO ALVARADO MOYA (FIRMA)
PERSONA FISICA, CPF-01-0753-0713.
Fecha declarada: 08/11/2022 08:09:36 AM
Razón: Acta de Aprobación Ronald Caravaca M.
Lugar: Cartago

Dr. Pablo Alvarado Moya
Profesor Lector

KERVIN SANCHEZ HERRERA (FIRMA)
PERSONA FISICA, CPF-02-0655-0231.
Fecha declarada: 08/11/2022 11:41:57 AM
Razón: Acta de Evaluación de Ronald Caravaca
Lugar: Alajuela

MSc. Kervin Sanchez Herrera
Profesor Lector

MARVIN COTO JIMENEZ (FIRMA)
PERSONA FISICA, CPF-01-1065-0799.
Fecha declarada: 08/11/2022 03:45:29 p. m.
Esta representación visual no es fuente
de confianza. Valide siempre la firma.

Dr. Marvin Coto Jiménez
Evaluador Independiente

LAURA CRISTINA
CABRERA
QUIROS (FIRMA)

Firmado digitalmente por
LAURA CRISTINA CABRERA
QUIROS (FIRMA)
Fecha: 2022.11.07 11:31:42
-06'00'

Dra. Laura Cabrera Quirós
Directora de Tesis

Cartago, Agosto 2022

Resumen

El procesamiento de señales sociales se ha desarrollado ampliamente en los últimos años, permitiendo el surgimiento de nuevas áreas de investigación, entre ellas, la predicción de atracción humana, donde se busca desarrollar sistemas computacionales que permitan un mejor entendimiento de la dinámica de la atracción. En esta tesis de investigación se plantea una metodología de aprendizaje de máquina multimodal que utiliza la sincronía de movimiento interpersonal para extraer patrones de comportamiento, que permitan predecir automáticamente la atracción humana en eventos de citas rápidas, ayudando a una mejor comprensión de esta. Se propone implementar un descriptor de movimiento para una de las modalidades (Video), una etapa de extracción de características de sincronía, un modelo de representación multimodal y un clasificador para la predicción. La evaluación de la metodología se hará comparando, mediante una prueba de hipótesis (Wilcoxon), el rendimiento del clasificador contra los resultado de estudios unimodales del estado del arte.

Palabras clave: Atracción humana, comportamiento humano, aprendizaje de máquina multimodal, auto-codificador multimodal, video, aceleración, representación multimodal, sincronía, movimiento interpersonal, predicción, Wilcoxon.

Abstract

The Social Signal Processing (SSP) has expanded widely in recent years, allowing the emergence of new areas of research, including the prediction of human attraction. The goal is to develop computational systems that give a better understanding of the dynamics of human attraction. In this research, a multimodal machine learning methodology is proposed, using interpersonal movement synchronization to extract behavior patterns, which automatically predict human attraction in speed dating events, helping to better understand it. We propose to implement a motion descriptor for one of the modalities (Video), a motion synchrony feature extraction stage, a multimodal representation model and a classifier for prediction. The evaluation of the methodology will be done by comparing, through a hypothesis test (Wilcoxon), the classifier performance against the results of unimodal research in the state-of-the-art.

Keywords: Human attraction, human behavior, multimodal machine learning, multimodal autoencoder, video, acceleration, multimodal representation, synchrony, interpersonal movement, prediction, Wilcoxon.

To my wife and my dear parents

Acknowledgments

First of all, I would like to thank God for the opportunity to finish this thesis and my master's studies. I would also like to thank my wife for her support, without which this work would not be possible. I would also like to thank my parents who have always been a great motivation in my career.

On the other hand, I would like to express my thanks to Dr. Laura Cabrera, who gave me the opportunity to work with her and has guided me in the development of the thesis. Finally, I would like to thank my colleague Carlos Brenes for working with me in the preparation of the data.

Ronald Gerardo Caravaca Mora

Cartago, 26 de agosto de 2022

Contents

List of Figures **iii**

List of Tables **v**

1 Introduction **1**

1.1 Context . . . 1

1.2 Problem . . . 2

1.3 Proposed solution . . . 3

1.3.1 Proposed hypothesis . . . 4

1.4 Objectives . . . 5

1.4.1 General Objective . . . 5

1.4.2 Specific Objectives . . . 5

1.5 Document structure . . . 5

2 Related work **7**

2.1 Previous Work . . . 7

2.2 Dataset . . . 8

2.2.1 Questionnaires . . . 9

2.2.2 Sensors . . . 9

2.2.3 Speed-dates statistics . . . 9

2.2.4 Balance of classes . . . 10

2.3 Synchrony of interpersonal movement . . . 10

2.3.1 Motion synchrony features . . . 11

2.4 Multimodal Machine Learning . . . 12

2.5 Multimodal Fusion . . . 13

2.5.1 Early fusion . . . 13

2.5.2 Late fusion . . . 13

2.6 Multimodal Representation . . . 14

2.6.1 Coordinated representation . . . 14

2.6.2 Joint representation . . . 15

2.6.3 Autoencoders . . . 15

2.6.4 Autoencoders for classification . . . 16

2.6.5 Multimodal Autoencoder . . . 17

2.7 Evaluation metrics . . . 18

2.7.1 Receiver operating characteristic (ROC) . . . 18
2.7.2 Precision-Recall Curve (PR curve) . . . 19

3 Proposed solution 21
3.1 Dataset preprocessing . . . 21
3.2 Extraction of motion synchrony features . . . 22
3.2.1 Acceleration modality . . . 23
3.2.2 Video modality . . . 23
3.3 Multimodal representation . . . 26
3.4 Classification . . . 27

4 Results and analysis 29
4.1 Reproducing the state-of-the-art . . . 29
4.1.1 Using baseline features . . . 29
4.1.2 Using motion synchrony features only . . . 31
4.1.3 Our proposed video baseline . . . 32
4.2 Multimodal representation . . . 34
4.3 Results of classification . . . 36
4.4 Further analysis . . . 38

5 Conclusions and future work 43
5.1 Conclusions . . . 43
5.2 Future work . . . 44

Bibliography 45

List of Figures

1.1 Proposed solution. . . . 4

2.1 Percentage of positive class of each problem by gender. . . . 10
2.2 Synchrony features extraction process [9]. . . . 12
2.3 Early fusion [1]. . . . 13
2.4 Late fusion [1]. . . . 14
2.5 Coordinated representation [3]. . . . 15
2.6 Joint representation [3]. . . . 15
2.7 Autoencoder model [23]. . . . 16
2.8 Autoencoders for classification [23]. . . . 17
2.9 An example of four ROC curves. [16] 18
2.10 An example of four PR curve 19

3.1 (a) MatchNMingle video frame and (b) data extraction with trackers. . . 22
3.2 Process to extract synchrony features from acceleration modality. . . . 23
3.3 Image grid for local optical flow processing: (a) Grid on real frame, (b) Grid on magnitude of the optical flow, (c) Plots of magnitude by grid cell, and (d) Plots of direction by grid cell. . . . 25
3.4 Discretizing angles in four directions with semantic meaning Front (F), Back (B), Right (R) and Left (L). . . . 26
3.5 Motion extraction from video process. . . . 26
3.6 Autoencoder for proposed solution. . . . 27

4.1 Autoencoder Training and validation Loss functions. . . . 34
4.2 Mean MSE and Standard Deviation for Autoencoder using testing dataset. 35
4.3 t-SNE visualizing classes in Autoencoder latent space. (a) SeeAgain problem and (b) Match_SeeAgain problem. . . . 36
4.4 (a) Average confusion matrix and (b) Precision-recall curve for SeeAgain individual interest using both genders. Positive to Negative class balance percentage is 58:42 (%). Positive class is 1. . . . 39
4.5 (a) Average confusion matrix and (b) Precision-recall curve for SeeAgain mutual interest using both genders. Positive to Negative class balance percentage is 36:64 (%). Positive class is 1. . . . 40

4.6 (a) Average confusion matrix and (b) Precision-recall curve for Romantic mutual interest using both genders. Positive to Negative class balance percentage is 15:85 (%). Positive class is 1. 40

4.7 Precision-recall curve for Romantic individual interest using using males only. Positive to Negative class balance percentage is 43:57 (%). Positive class is 1. 41

4.8 (a) Average confusion matrix and (b) Precision-recall curve for Friendly mutual interest using using males only. Positive to Negative class balance percentage is 27:73 (%). Positive class is 1. 41

List of Tables

3.1 Percentage obtained from videos and acceleration data. 22

3.2 Video modality position-base features [36]. 24

4.1 Mean AUC±STD results with variance in distance with video modality using MnM dataset. 30

4.2 Mean AUC±STD results with position variance with video modality using MnM dataset. 30

4.3 Mean AUC±STD results with variance in the other person's position with the video modality using MnM dataset. 30

4.4 Mean AUC±STD results with the acceleration mode for individual interest using MnM dataset. 31

4.5 Mean AUC±STD results with the acceleration modality for mutual interest using MnM dataset. 31

4.6 Mean AUC±STD results for both modalities with synchrony features with male and female data using MnM dataset. 32

4.7 Mean AUC±STD results for both modalities with synchrony features and only male data using MnM dataset. 32

4.8 Mean AUC±STD results for both modalities with synchrony features and only female data using MnM dataset. 33

4.9 Mean AUC±STD results for comparison of state-of-the-art baseline (STD B) and our proposal (OUR B). $*$ indicates that the result significantly outperforms the video baseline with $p < 0.05$. 33

4.10 Mean AUC±STD results *Majority Voting* classifier with synchrony features using MnM dataset. 34

4.11 Mean AUC±STD for unimodal and multimodal approaches using *both genders*, synchrony features and MnM dataset. $*$ indicates that the result significantly outperforms the video baseline with $p < 0.05$, $\circledast$ indicates that the result significantly outperforms the acceleration baseline with $p < 0.05$ and † indicates that the result significantly outperforms both baselines with $p < 0.05$. 36

4.12 Mean AUC±STD for unimodal and multimodal approaches using *males* only, synchrony features and MnM dataset. $*$ indicates that the result significantly outperforms the video baseline with $p < 0.05$, $\circledast$ indicates that the result significantly outperforms the acceleration baseline with $p < 0.05$ and † indicates that the result significantly outperforms both baselines with $p < 0.05$. 37

4.13 Mean AUC±STD for unimodal and multimodal approaches using *females* only, synchrony features and MnM dataset. $*$ indicates that the result significantly outperforms the video baseline with $p < 0.05$, $\circledast$ indicates that the result significantly outperforms the acceleration baseline with $p < 0.05$ and † indicates that the result significantly outperforms both baselines with $p < 0.05$. 38

Chapter 1

Introduction

1.1 Context

The development of computers, robotics and mobile devices enabled significant progress in the development of interfaces that allow human-computer interaction (HCI). The recognition of voice, gestures, emotions and body movements through these types of interfaces has been highly exploited by today's society. This has allowed human-computer interaction to go beyond just text, and opened up a new field of research called *Social Signal Processing*. This field is defined by Vinciarelli [38] as *"the computing domain aimed at modeling, analysis, and synthesis of social signals in human–human and human–machine interactions"*.

As Vinciarelli et al. [39] states, the social signals are manifested through voice inflections and multiple nonverbal behaviors such as facial expressions, gestures and postures. It also explores how social behaviors, such as empathy, friendliness, and attraction, are expressed as temporal patterns of specific social behavioral cues during a social interaction. In addition, Vinciarelli et al. [39] states that these behaviors are not aleatory and could determine the behavior of other people in a social interaction or in the context of it.

According to Vijayalakshmi et al. [37], one of the fields where social behaviors have been studied through social signals is the *emotional field*, using facial expressions, voice inflections and gestures, in order to analyze certain behaviors where emotions play an important role.

Among other areas of research in this line, we have the detection and prediction of human attraction. According to Veenstra et al. [36], finding a partner may not be complicated for many people, however, a large number of them face various problems in finding a person suitable to their preferences. This is why Internet dating sites, online pages that help find a partner or search for people who are in some way compatible, are on the rise. According to an article published by teletica.com [31], some experts believe that around the world by 2030 Internet dating sites will be much more common than they are today, making it a growing industry and a developing area of research.

Motivated by the above, this thesis aims to automatically predict human attraction using video signals and motion acceleration, to understand whether certain non-verbal behaviors demonstrate romantic interest or not during social interactions. The importance of automatic prediction lies in the need to understand the dynamics of human attraction to somehow model this behavior, making machines capable to process and understand human attraction effectively. It is also important to develop attraction detection systems capable of using human social cues to support people in finding a partner. In other words, to make people have a better experience and comfort when faced with the problem of finding a couple.

Existing research about automatic detection of human attraction implements a unimodal approach, i.e. using a single type of data such as acceleration, video or audio. The main limitation of these methods is that they can not identify specific characteristics that could be achieved with another modality. In other words, do not have a complete picture of the event, only a single perspective that may be insufficient to determine the presence of attraction. For example, with acceleration sensors, general movements of the body as a whole could be detected, as well as fine movements and axial movements; but exclusive movements of the head or limbs might not be captured. With video, movements of limbs and specific parts of the body could be more easily captured. Other relevant cues such as position, distances and angles with respect to a reference could be determined, but axial movement data could not be obtained as with accelerometers.

Therefore, this research thesis attempts to take advantage of the multimodal nature of social interactions by capturing data from two different modalities that have two different perspectives of the same event. Specifically tri-axial acceleration captured with accelerometers and the video of people during the social interaction. With the information from these two modalities we propose to make a multimodal representation by training a Machine Learning model, which uses the motion information from both modalities (perspectives), to improve performance in predicting human attraction in social interactions. Such social interaction will be taken from speed-dates, which are events where people meet potential romantic partners for the first time during short intervals of time in the range of minutes and then indicate whether they want to meet the partner again.

1.2 Problem

In order to gather information to understand the key components of human attraction, much of the related research has been conducted through surveys, interviews, and form-filling. However, these methods take a long time to analyze, the data processing is tedious, do not provide a prompt response, are clearly invasive, and directly interfere with people's interactions. But, in reality the process of human attraction occurs in interactions where there is usually no interference from third parties, people want to feel at ease in a safe and undisturbed environment, where they can interact freely in total comfort and know as soon as possible whether they were successful or not in the speed date. Consequently, there is a need to understand the dynamics of human attraction in social interactions without interfering with them, promptly and in a real environment. Therefore, there is a need to assess dyadic interactions in a non intrusive manner to automatically determine if there is attraction between two people or not without the intervention of third parties

1.3 Proposed solution

In the field of psychology, research has been done to determine whether a behavior reflects attraction or not. As in the work done by Guéguen [14], Karremans et al. [22] and Farley [11], who found that mimicry and synchrony in movement have a positive correlation with romantic interest. Also, the work of Nanninga et al. [26] and Ramseyer et al. [28] suggests that synchrony, head movement and coordination are factors that can predict attraction.

This premise of using motion synchrony has already been used in other research, such as in the work of Kapcak et al. [21] and Veenstra et al. [36]. Here they used, among others, synchrony features to detect human attraction in a social interaction. However, they employed data from only one modality at the time, acceleration and video respectively.

Taking these works and those from social psychology as motivation for this thesis, we have that nonverbal behavior reflected in body movement is a key indicator of romantic interest. This is why we intend to use video and acceleration modalities to capture nonverbal behaviors associated with movement to analyze them in a multimodal approach, by extracting features of motion synchrony through the acceleration in the three axial axes and the movement present in the video. We then use these to assess inter and intra-personal synchrony movement. With this, we aim to take advantage of the different perspectives that each modality can give of the same event and the possible correlation between data from different sources.

This data needs to be processed to find patterns that determine whether there is attraction or not. However, discovering these patterns is a complex task for the average person and requires a lot of analysis time. This is why using a Multimodal Machine Learning model is so important for this development. As shown by [36, 29] and [21], Machine Learning models can even improve attractiveness prediction over human perception by up to 15.3% in men and by 6.8% in women. This is because these models could find patterns in large amounts of data that could not be detected by an ordinary person. Also, Machine Learning could automate all the process to requires less analysis time.

Our proposed solution is illustrated in Figure 1.1. From two modalities, motion synchrony features are obtained, which we call, *low-level features*. In the specific case of acceleration, these features can be obtained directly from accelerometer data in the three axial axes, while for video, a stage where the motion can be quantified is necessary. For this purpose, it is proposed to develop a method to extract the motion present in video from which the motion synchrony features are obtained.

The synchrony features from both modalities are concatenated and used as input to a

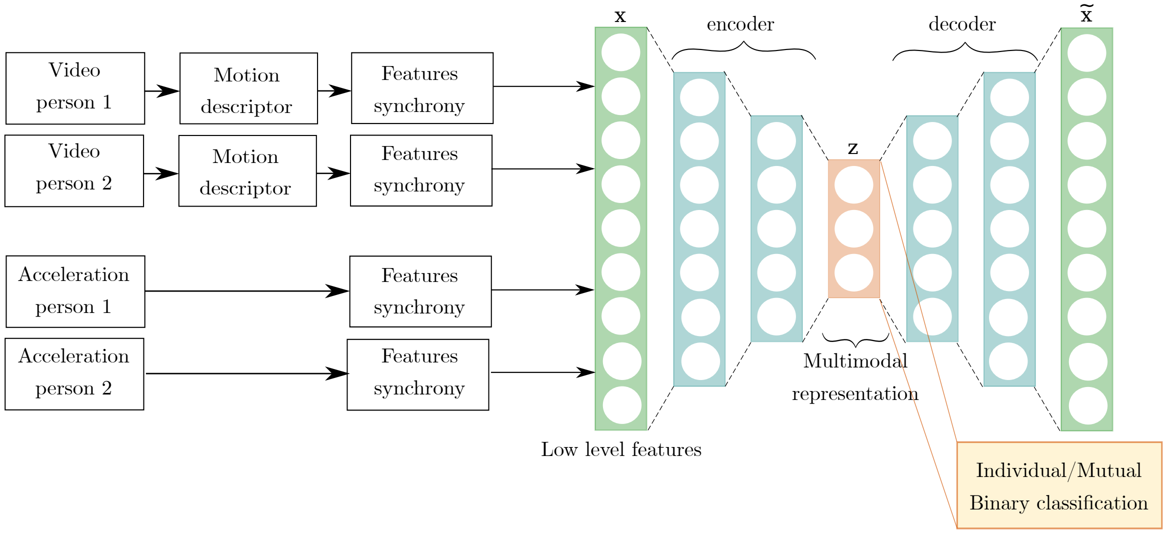


Figure 1.1: Proposed solution.

autoencoder (a neural network that learns a reduced data representation ignoring insignificant data to reconstruct the input), in order to obtain a multimodal representation of the synchrony, which is also reduced in dimension. This representation space, we will call it *Multimodal Representation*.

Afterwards, the representation space obtained is then used as input to a classifier, to evaluate two cases of attraction. The first one, **individual interest case**, i. e. whether one of the participants was attracted to the other. Second one, **the mutual interest case**, i. e. whether both participants were attracted to each other. For each case, we have four binary classification problems that represent a different type of attraction interest, each of them could have a specific nonverbal behavior so we will study the four problems to understand their implications. The four classification problems are: *SeeAgain, Friendly, Sexual* and *Romantic*, when it concerns to individual interest, and these four *Match_SeeAgain, Match_Friendly, Match_Sexual* and *Match_Romantic* when it concerns to mutual interest. Although we will study all four problems, we will use *SeeAgain* as the main case of study.

1.3.1 Proposed hypothesis

A statistically significant improvement in detection performance over unimodal baselines in the prediction of human attraction will be obtained if motion synchrony is leveraged from two different modalities.

1.4 Objectives

This section presents the general objective and the specific objectives to be covered in the development of the thesis, in order to provide a clearer picture of the scope and limits of the thesis.

1.4.1 General Objective

Develop a multimodal method for automatic prediction of human attraction using video and wearable acceleration sensors.

1.4.2 Specific Objectives

1. Adapt a multimodal dataset to specific requirements of this research.
2. Propose and implement feasible unimodal baselines for video and acceleration modalities.
3. Implement the multimodal representation model with an autoencoder of the proposed solution.
4. Evaluate whether the proposed solution provides a statistically significant improvement over the state of the art.

1.5 Document structure

The following is the structure of the document containing the information and the steps to be followed to achieve the objectives of this research. The concepts that provide the theoretical basis for the proposed solution are presented in Chapter 2, the proposed solution is developed in Chapter 3, the results and analysis are described in Chapter 4, and finally in the Chapter 5 we present our conclusions and future work.

Chapter 2

Related work

In this section we describe the theoretical concepts that support the methodology to be developed. We present the bibliographical study of previous work, dataset, features for each modality according to the state of the art and concepts of multimodal machine learning.

2.1 Previous Work

One of the first research in the field of Automatically detecting human attraction was conducted by Madan et al. [24], who studied romantic, friendly, and commercial interest between people through audio analysis, extracting four main elements: activity, engagement, emphasis, and reflection. Using linear classifiers, they achieved a 71% of accuracy. Another audio-based research was realized by Michalsky et al. [25] who identified that attraction behaviors can be detected through the tone of voice. Ranganath et al. [29] also studied the difference between intention and perception through speech in speed-dates, where both the speaker and the listener rated the speaker as someone who was "flirting". Their "flirting" detection system uses prosodic, dialogue and logic functions to detect whether a speaker is trying to flirt, with an accuracy of up to 71.5%.

In previous research, the audio used was recorded at speed-dating events using portable recorders and microphones, which participants wore close to their faces. These research have shown interesting results, however, from the point of view of privacy, audio is not the best modality, since it requires recording people's own words, invading their intimacy. Some people reported to not feel completely free [21] to express themselves and may also experience some discomfort knowing that their words are being recorded, generating inaccurate data as it is unclear whether a certain voice intonation is due to attraction or discomfort.

More recent research uses less intrusive modalities such as overhead video recordings and wearable acceleration sensors in speed-dates. In the case of video, trajectory and position tracking methods are used to detect whether certain movements or nonverbal behaviors are evidence of physical and sexual attraction. Veenstra et al. [36] used videos of eight women and eight men in 64 speed-dates events, with an average age of 23.4 years, and cameras placed on top of the participants. They used location and relative position of the participants as features of attraction, that were extracted through a combination of *background subtraction* and *clustering*. The cluster *centroids* represent the location of the participants which are tracked to obtain the trajectories and position. Using SVM and KNN classifiers they predicted attraction as a binary problem with an accuracy of 70% for females and 72% for males for certain motion characteristics, whereas using only synchrony-base features they achieved an prediction accuracy of 48% in males and 55% in females. This research is the state of the art for automatic attraction detection using the video modality.

Using wearable devices attached to people's chest like a badge, Kapcak et al. [21], developed a method that consists in modeling the nonverbal behavior extracted from a tri-axial accelerometer. The mean, variance, and power spectral density of each of the axes are extracted. From these, they obtain features such as correlation, mutual information, metrics, and convergence between the signals of both persons achieving an AUC of 80%. They performed the evaluation using 10-fold cross-validation. However, this approach could overfit the model because there could be information from the same participant in the training and test dataset. It is important to mention that there is an extended version of this research published by Vargas Quiros et al. [35]. This research is the state of the art for the acceleration modality.

It should be noted that both Veenstra et al. [36] and Kapcak et al. [21] used synchrony-based features but in a unimodal approach.

To the best of our knowledge, there is no development that implements a multimodal approach with speed-dates data to detect human attraction using video and acceleration. Therefore, compare to previous work in this thesis we further explore how multimodal analysis impacts the detection of nonverbal behaviors of human attraction, and whether somehow combining multiple modalities improves the detection performance.

2.2 Dataset

The dataset used in this work is *MatchNMingle* proposed by Cabrera-Quiros et al. [6]. It is a multimodal dataset created specifically to contribute to the automatic analysis of social signals and interactions. It consists of 4 hours of uninterrupted recording of conversations of 92 people. The data were taken in a real-life setting, during three speed dating events on three different days.

The data collected include acceleration and video, using wearable accelerometers in a badge-like device hung around the neck and overhead cameras. The participants were single, heterosexual, college students between the ages of 18 and 30 years old. There were 30 people per event with 15 males and 15 females. The available data are as follows:

2.2.1 Questionnaires

1. *Speed-dates responses:* After each speed-dates session, participants indicated whether they were interested in the person with which they just interacted by answering the following questions using a 7-point Liker scale, with 1 being the lowest and 7 being the highest:

   - How much would you like to see this person again?
   - How would you rate this person as a potential friend?
   - How would you rate this person as a short term sexual partner?
   - How would you rate this person as a long term romantic partner?

   Each of these questions is used as *ground truth* for four different classification problems; *SeeAgain, Friendly, Sexual and Romantic*, when it comes to individual interest. And these four; *Match_SeeAgain, Match_Friendly, Match_Sexual, Match_Romantic* when it comes to mutual interest.

2. *Frontal photography:* In each event, frontal photographs were taken of each participant with different facial expressions: neutral, smile, and full body. These photographs are necessary in the preparation of the data to identify the videos corresponding to each person.

2.2.2 Sensors

Two modalities were used to collect nonverbal behavioral data from the participants:

1. *Accelerometers:* Each participant was given a sensor in the form of a badge. These sensors were designed especially for social events. The sensors collect tri-axial data at 20 Hz.

2. *Video cameras:* These data were collected using cameras with a resolution of $1920x1080$ (16:9), at a sampling rate of 30 FPS in a wide vision. They are located at the top of the room to avoid obstructions from the participants themselves.

2.2.3 Speed-dates statistics

In accordance with Cabrera-Quiros et al. [6], a *date* refers to the information of one participant during a 3-minute date, whereas a *date-interaction* refers to the interaction of two people during a 3-minute date.

Which means that each participant had a *date-interaction* with each participant of the opposite sex. Thus, for day 1 each participant had about 14 dates and for days 2 and 3 about 15 dates. In total 674 *date-interactions* were collected. However, due to some sensors did not work properly, we did not have data from some participants, so we managed to obtain only 65% of *the interactions*, in other words 435 out of 674 (Day 1 = 224, Day 2 = 225, Day 3 = 225).

It should be noted there are *date interactions* that have data from one participant only, due to problems in some sensors and as a result of data processing.

2.2.4 Balance of classes

In order to make a better result of the analysis it is important to know the percentage of positive classes by classification problem we have, after a binarization process described in 3.4. In Figure 2.1 we can see how the positive classes are distributed by classification problem. We can observe that for individual interests there is a good balance between positive and negative classes, but for mutual interests there is a clear unbalance by having much less positive classes than negative. This unbalance could affect the performance of the classifier.

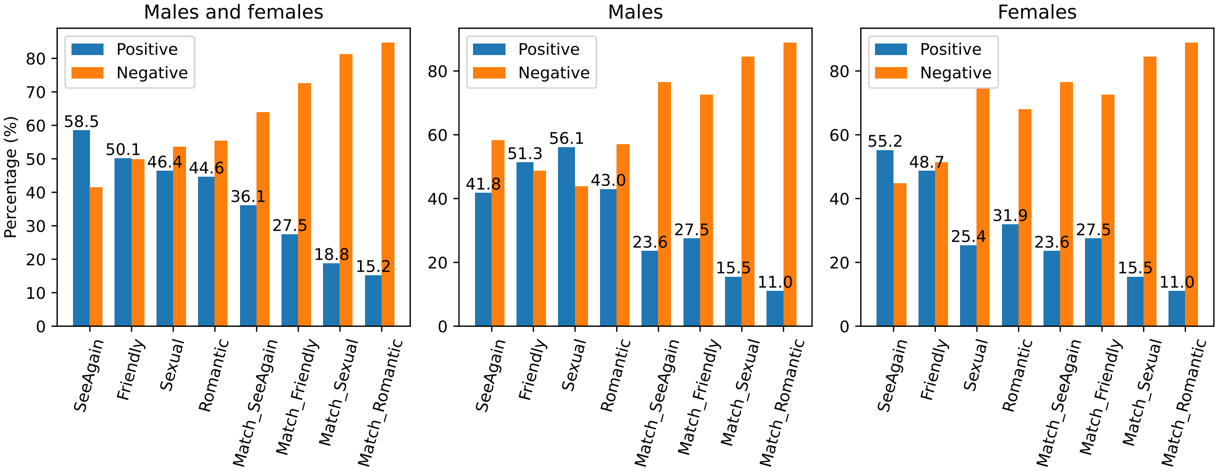


Figure 2.1: Percentage of positive class of each problem by gender.

2.3 Synchrony of interpersonal movement

As described in Section 1.3, we will use motion synchrony features in order to capture non-verbal behavior patterns. These features are based on works about Synchrony of interpersonal movement that we describe as follow.

Chetouani et al. [7] define synchrony of interpersonal movement or *motion synchrony*, as the dynamic and reciprocal adaptation of the temporal structure of behaviors between interacting people. Unlike reflection or mimicry, synchrony is dynamic in the sense that the important element is synchronization rather than the nature of the behaviors. However, the distinction between synchrony and reflection can be blurred. These phenomena are not mutually exclusive and can usually be observed simultaneously.

Also, Chetouani et al. [7] and Delaherche et al. [9] propose to use automatic techniques to capture social cues and evaluate motion synchrony in human-human interactions.

The first step in calculating synchrony is to extract the relevant motion features of the dyads with motion tracking devices, image processing techniques or physiological sensors. After extracting the motion features, a similarity measure is applied. *Correlation* is the most used method to assess interpersonal synchrony. Also, a *time-delayed cross-correlation* is applied between the movement time series of the dyads by means of short interaction windows.

Another method for evaluating the similarity of movement of two people is *recurrence analysis*. This evaluates the points in time at which two systems show similar patterns of change or movement, called "recurrence points".

There are also *spectral methods*, which are an alternative to temporal methods for rhythmic tasks. Spectral methods measure the relative phase evolution between the two persons. They also measure the superposition between the movement frequencies of the two, called *cross-spectral coherence* or *power spectrum superposition*.

2.3.1 Motion synchrony features

The Motion synchrony features must be extracted from statistical data such as mean, variance and power spectral density obtained from signals divided into temporal windows. From these data we can extract the features describes as follow:

1. Correlation: It is used to measure the similarity of body movement and body parts such as hands and heads of two people, as indicated by Ramseyer et al. [28] and Tschacher et al. [33], through the Pearson Correlation of Equation 2.1,

$$\rho_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sigma(X)\sigma(Y)} \tag{2.1}$$

   where x and y are two distributions of N values that represent the movement information of two persons. μ_x and μ_y are the mean values of x and y distribution respectively, and $\sigma(X)$ and $\sigma(Y)$ are the variance of x and y distribution respectively. A value equal to 1 is expected when two people show interest in each other.

2. Mutual information: It is used to calculate the dependence between who interacts. This is shown by Cabrera-Quiros et al. [5] and Gedik et al. [12]. In this case, dependence between two people is captured, and it is calculated as:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{2.2}$$

where, $H(X)$ and $H(Y)$ represent the entropy of the random variables X and Y and $H(X, Y)$ the joint entropy. In addition, the information is normalized and values between 0 and 1 are obtained.

3. Mimicry: The objective here is to determine when one person mimics the other. In this case, each sample of person A is compared with the consecutive sample of person B, obtaining the distance between the low-level features, and using the minimum, maximum, mean and variance as features.

4. Time-lagged correlation: This is used to measure the correlation between the movement of the person following the conversation and the movement of the person speaking. The correlation formula:

$$\rho_{xy} = \frac{\sum_{i=1}^{N-\tau}(x_i - \mu_x)(y_{i+\tau} - \mu_y)}{\sigma(X)\sigma(Y)} \tag{2.3}$$

is used for this purpose, where x and y are two distributions of N values that represent the movement information of two persons. μ_x and μ_y are the mean values of x and y distribution respectively, $\sigma(X)$ and $\sigma(Y)$ are the variance of x and y distribution respectively and τ represent a shift in time. A value equal to 1 is expected when two people show interest in each other.

This process of extracting the synchrony features is illustrated in Figure 2.2, where the motion of each person is extracted and then the individual and dyad features are extracted.

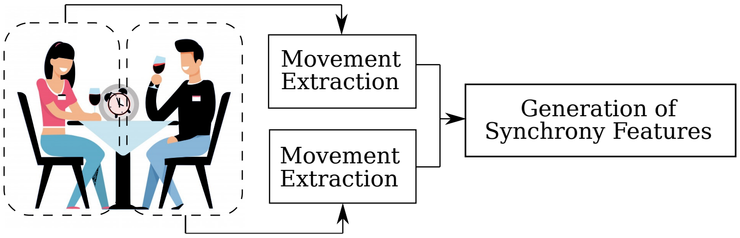


Figure 2.2: Synchrony features extraction process [9].

These types of features can be found in a single modality as shown in [21, 36], or in different modalities as summarized by Chetouani et al. [7].

2.4 Multimodal Machine Learning

According to Baltrusaitis et al. [3], *multimodal machine learning* seeks to create models that can process and relate data coming from different modalities, i.e., coming from different sources or measured using different types of signals, giving the possibility to obtain correspondences between them and give a better understanding of the problem. Baltrusaitis et al. [3] identify five main technical challenges for the implementation of multimodal machine learning, which are: representation, translation, alignment, fusion, and co-learning. Of these fice challenges, the one that concerns this research is *representation*, which consists in how to represent multimodal data in a way that takes advantage of the redundancy and complementarity of multiple modalities. Therefore, the solution proposed in this thesis aims to represent the motion synchrony characteristics in a multimodal space, where correlation between video and acceleration modalities are leveraged. (See Section 1.3)

2.5 Multimodal Fusion

Atrey et al. [1] state that within Multimodal Machine Learning, the integration of multiple modalities, their features or intermediate decisions to perform an analysis task is called multimodal fusion. The fusion of different modalities is generally performed at two levels: *early fusion* and *late fusion*.

2.5.1 Early fusion

In this strategy there are classifiers or analysis units AU that provide the D_n decisions based on the F_n features of each individual modality. The decisions are merged into a decision vector that is further analyzed to obtain the final decision, as shown in Figure 2.3.

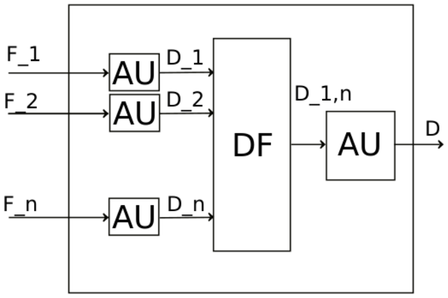


Figure 2.3: Early fusion [1].

Early fusion has the advantage of using correlation between different modalities at an early stage which helps to improve the analysis. In addition, it requires only one learning phase.

2.5.2 Late fusion

In this case the features are extracted from the data, combined and used as input to a single classifier or analysis unit AU. For example, in a face detection model, multimodal features such as skin color and motion signals are combined into a single larger feature vector that is taken as the input to the detection model. Figure 2.4 describes this process, where F_n represents the features, FF the feature fusion unit and AU the analysis unit. An example of this late fusion, is the *Majority Voting* method, where the final decision is the one in which the majority of classifiers reach a similar decision.

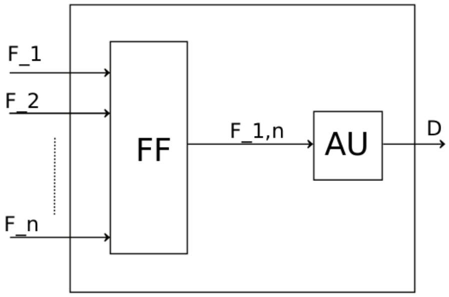


Figure 2.4: Late fusion [1].

2.6 Multimodal Representation

As Baltrusaitis et al. [3] explain, an entity can be represented by vectors or tensors that contain the data of this entity, for example, entities such as images, audio, words, or sentences. A multimodal representation includes data that contains information from multiple entities. This presents several challenges, such as combining data from heterogeneous sources, handling different types and levels of noise, and loss of data from some of the modalities.

Bengio et al. [4] identify properties that make a good representation: smoothness, temporal and spatial coherence, sparsity, and natural grouping, among others, but mainly it is required that the representation has to be easy to obtain even in the absence of some modalities, and it must be possible to complete the missing modalities given the observed ones.

For a better understanding of multimodal representation, in [3] it is proposed to divide it into two categories: *coordinated representation* and *joint representation*.

2.6.1 Coordinated representation

In the coordinate representation, each unimodal signal is processed separately but requires similarity constraints on each one to bring them into what is called a coordinate space, as shown in Figure 2.5. Mathematically it is described as:

$$f(\boldsymbol{x}_1) \sim g(\boldsymbol{x}_2) \tag{2.4}$$

where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are the input vectors by modality, and (f, g) are the functions that project each modality into the coordinate space. Examples of these functions are minimizing the cosine distance or maximizing the correlation.

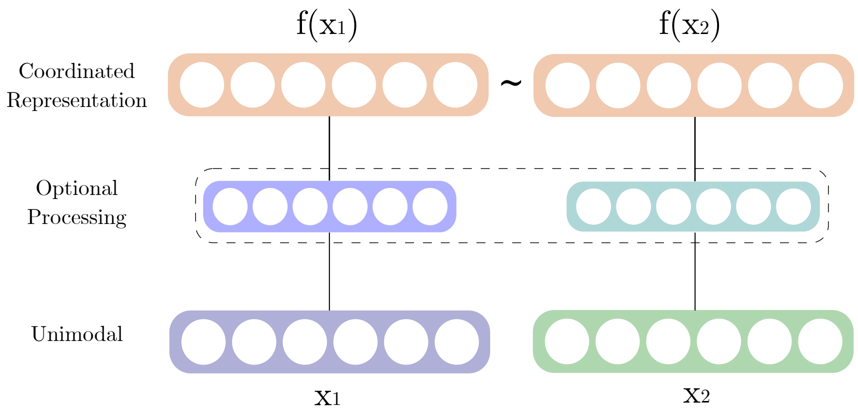


Figure 2.5: Coordinated representation [3].

2.6.2 Joint representation

The joint representations combine the unimodal signals in the same representation space, as shown in Figure 2.6. Mathematically it is described as:

$$\boldsymbol{x}_m = f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_2) \tag{2.5}$$

where the multimodal representation $\boldsymbol{x}_m$ is obtained by means of the function f, which can be a Neural Network, CNN, RNN, Boltzmann Machine, Autoencoders, etc.

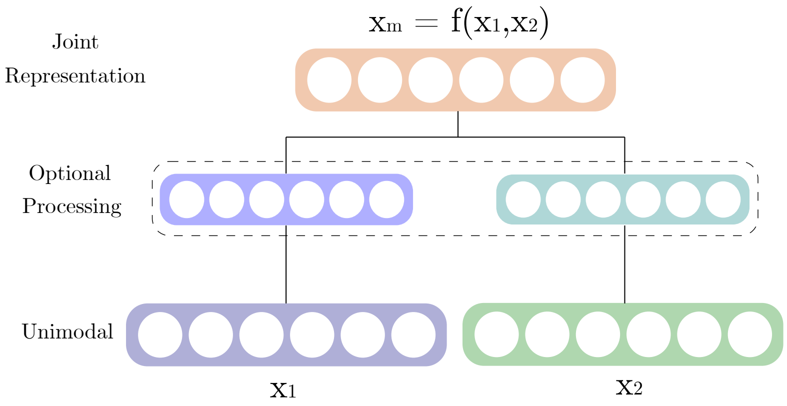


Figure 2.6: Joint representation [3].

2.6.3 Autoencoders

For the purpose of this research, we will use a joint representation method called *autoencoders* [3]. They could be defined as early or late fusion depending on requirements. According to Lopez Pinaya et al. [23], autoencoders are neural networks trained to reconstruct an input in an unsupervised way, removing unnecessary information from the input and taking advantage of any possible relations between the data required to reconstruct the entry.

The main function of the autoencoders is to learn a useful representation of the data, usually in a compressed format, that is performed by a neural network called an *encoder*. Then this representation is reconstructed by another neural network called *decoder* [18]. Thus, the problem consists of learning the functions $A : \mathbb{R}^n \to \mathbb{R}^m$ (encoder) and $B : \mathbb{R}^m \to \mathbb{R}^n$ (decoder), where $m < n$, satisfying [2]

$$\arg\ \min_{A,B} E\left[\Delta(\boldsymbol{x}, B \circ A(\boldsymbol{x}))\right] \tag{2.6}$$

where E represents the expectation over the distribution $\boldsymbol{x}$, and Δ the reconstruction loss function, which is usually obtained by measuring the distance between the decoder output and the encoder input, using $\ell_2 - norm$ and, A and B being neural networks [23].

The autoencoders generate a *latent space*, also called *latent representation*, or *reduced dimensional representation space* $\boldsymbol{z}$ similar to what is achieved with Principal Component Analysis (PCA), where a low-dimensional hyperplane is obtained in which the data is located, so it can be said that an autoencoder is a generalization of Principal Component Analysis [27]. This dimensionality reduction is obtained by imposing a "bottleneck" by reducing the size of the hidden layers of the encoder neural network. This representation can be used for purposes such as data compression, noise reduction and feature extraction. Figure 2.7 shows an illustration of the model, where $\boldsymbol{x}$ represents the input data, $\boldsymbol{z}$ the latent representation and $\tilde{\boldsymbol{x}}$ the output of the decoder, i.e. the reconstructed input.

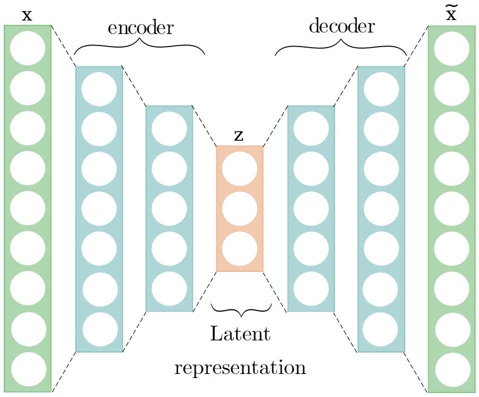


Figure 2.7: Autoencoder model [23].

2.6.4 Autoencoders for classification

While autoencoders are trained in an unsupervised manner, they can also be used in a semi-supervised methodology, where part of the data is labeled, to improve the results of a classification. In this case, the encoder is used as a feature generator in the latent space which is then taken as input to a classifier, as shown in Figure 2.8. The expectation is that samples with certain labels correspond to some specific latent representation.

First, the autoencoders are trained in an unsupervised way to obtain the latent representation. Then, the decoder is set aside and the encoder is used as input to a classification model, for example an SVM or some neuronal model.

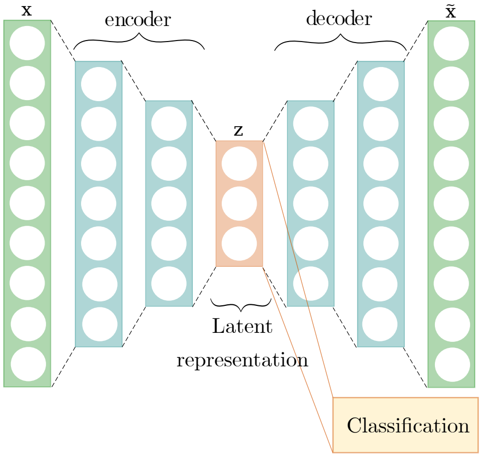


Figure 2.8: Autoencoders for classification [23].

2.6.5 Multimodal Autoencoder

According to Ghosh et al. [13] and Hu et al. [15], the autoencoder model is one of the most popular multimodal data fusion model, called *Multimodal Autoencoder*. In this model the input data consists of multimodal data, where each sample is denoted as $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$, having M modalities and the training set consists of N multimodal samples. The features of all modalities are concatenated in a single input vector and passed through the encoder stage, which mixes the information obtaining a latent multimodal representation; $\boldsymbol{z}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M)$, where it is expected to preserve the cross-modal semantic relations [8].

The main advantages of this multimodal representation model are the following:

1. The model can take advantage of correlation, redundancy, and interactions between the features of each modality. That is, it can extract relationships between modalities and learn how data from one modality to another complement each other.

2. The correlations between modalities can be used to recognize patterns that exist only between modalities.

3. Only one stage of unsupervised training is required, which facilitates the entire training process.

4. It is robust to noise and the absence of any of the modality.

2.7 Evaluation metrics

2.7.1 Receiver operating characteristic (ROC)

According to Huang et al. [16], the receiver operating characteristic (ROC) curve has been used in signal detection theory to represent the trade-off between hit rates and false hit rates. In recent years, it has been used to evaluate the performance of machine learning algorithms, using the relationship between false positive and true positive rates. That is, having the true positive (TP) and false positive (FN) rates predicted by a classifier, defined as:

$$TP = \frac{\text{positives correctly classified}}{\text{total positives}} \tag{2.7}$$

$$FP = \frac{\text{negatives incorrectly classified}}{\text{total negatives}} \tag{2.8}$$

We can create a graph with FP in x axis and TP in y axis, resulting in a curve called *ROC curve*. Figure 2.9 represents the ROC curves of 4 classifiers. Each curve represents one classifier. It is said that one classifier *dominates* another one, when its curve is above and to the left of the other curve. For example, in Figure 2.9 classifier A has a better performance than B and D. However, there are some cases where this is not so straightforward, for example, if we compare curves B and C we cannot determine which curve is above or to the left of the other. In these situations, the area under the ROC curve (AUC) is a convenient way to summarize and compare the ROC curves.

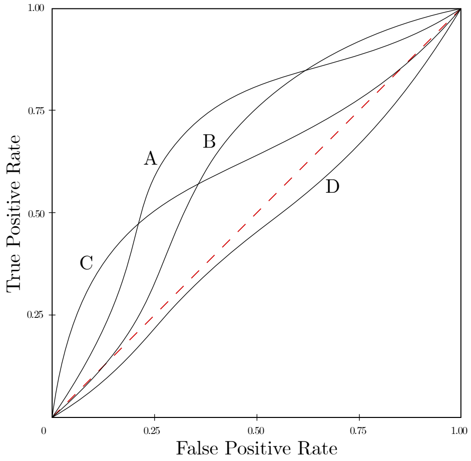


Figure 2.9: An example of four ROC curves. [16]

In accordance with Janssens et al. [17], the ROC is a probability curve and AUC represents a measure of separability. It indicates the capability of the model to discriminate between classes. An AUC close to 1 means that the model has a good class separability, when the AUC is 0.5 or less, it means that the model has no class separability at all.

2.7.2 Precision-Recall Curve (PR curve)

In classification problems, performance is usually defined by the confusion matrix generated by the classifier, from which it is possible to calculate precision and recall [19], defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2.9}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.10}$$

where TP are the number of True Positives, FP the False Positivse and FN are the False Negatives. Using these values we can plot a curve like Figure 2.10. To create the PR curve we need to vary the *decision threshold*. This figure shows an example of this curve for three different models, each with a different predictive capacity. Model A, which approaches the upper right corner, represents a high performing classifier as it has a good balance between precision and recall; the area under the curve is close to 1. The curves further away from that corner would represent lower performing models, such as model B. In the case of model C, this means that it is predicting the negative class in most cases.

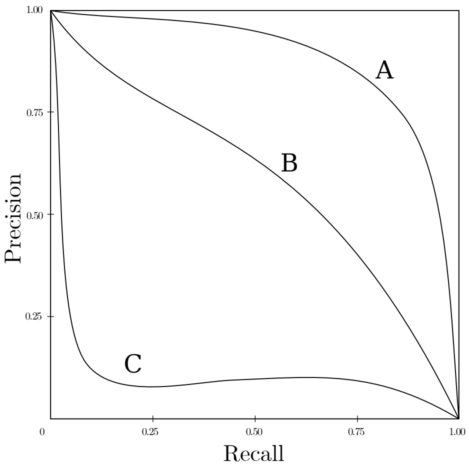


Figure 2.10: An example of four PR curve

We chose these metrics because we have an unbalanced dataset as shown in section 2.2.4. Both AUC and PR curves are robust to these unbalances. Metrics such as accuracy are affected by unbalanced datasets.

Chapter 3

Proposed solution

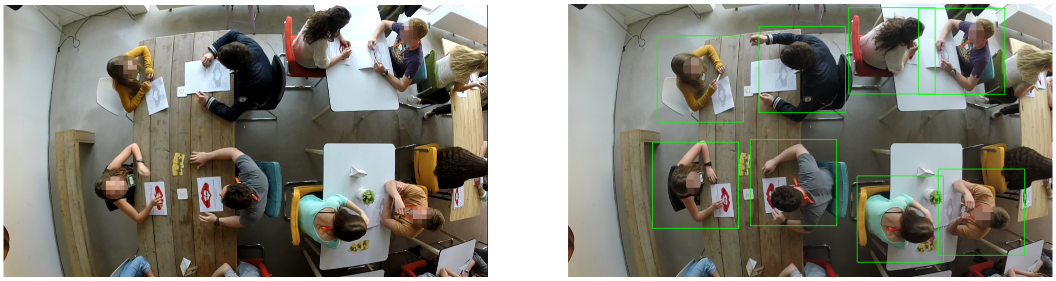
This chapter details the implementation of the proposed solution, as shown in Figure 1.3. Firstly we need to properly extract the low-level motion synchrony features for both modalities. These features contain the nonverbal motion information that we want to represent in the multimodal space to be used for classification problems.

In the case of the video modality, it is necessary to propose and implement a motion descriptor from which the low-level features are extracted. This is because the features used in the baseline are based on distances and position, they are not about motion synchrony. Also, in our dataset people are always sitting, they never walk or change their position with respect to the other person, so that we need to extract the motion from the video to obtain the synchrony features to use them as input of the autoencoder of the solution, to have the multimodal representation and then make the classification.

3.1 Dataset preprocessing

The MatchNMingle video data needs some preprocessing to be used in the proposed solution. The cameras recorded scenes with multiple people, as shown in Figure 3.1a. However, it is necessary to have an individual video for each person, so a system that extracts the videos of each person in a delimited area is needed. In addition, the relative locations of the people are needed, in order to extract the features.

The goal is to obtain the relative position of each person and delimit the area in which he/she is moving in order to create an individual video for each participant. So we proceeded to implement a system based on *OpenCV Motion Trackers*, specifically *Kernelized Correlation Filter (KCF)*. With this, the position of the object is extracted and the location of the center of the participants in the (x,y) plane is obtained, thus generating the displacement data for each person. The position is updated every 15 frames and the intermediate points are interpolated. Also, once the position is obtained, the area where it is located is delimited and the videos corresponding to each participant and their respective partner are generated, as shown in Figure 3.1b. In average, an 80% of the 1380 possible videos were obtained in the three events.



(a)

(b)

Figure 3.1: (a) MatchNMingle video frame and (b) data extraction with trackers.

In the case of the acceleration modality, we have tri-axial data and the elapsed time for each person during the entire event. Using the elapsed time we extract the data corresponding to each speed date.

Due to the problems described in Section 2.2.3, and due to specific problems of videos, it is not possible to obtain 100% of the data, both from the videos and from the accelerometers. The problems are as follows: people going out of frame in the original videos, errors in the time corresponding to each date during the event, and occlusions by people. Also, some wearable devices stopped working. Due to these problems, it is not possible to obtain all possible data from the three events. Table 3.1 shows the percentages of videos and acceleration data for both individuals and pairs. In other words, the number of dates containing information from at least one participant (dates) and the number of dates for which there is information from both participants (interactions). While there is a decrease in the amount of data, we still have enough data for training and testing.

Tabla 3.1: Percentage obtained from videos and acceleration data.

	Modality	**Day 1 (%)**	**Day 2 (%)**	**Day 3 (%)**
Dates	Video	92.92	87.78	60.89
	Acceleration	87.50	61.33	76.67
	Video ∩ Acceleration	86.46	61.33	48.44
Interactions	Video	96.88	76.89	60.89
	Acceleration	82.14	50.22	53.78
	Video ∩ Acceleration	79.02	38.22	32.00

3.2 Extraction of motion synchrony features

In order to have the motion synchrony features from video and acceleration modalities which describe the non verbal behavior present in speed dates, the features introduced in Section 2.3 will be used. They are related to *correlations, recurrence and spectral characteristics* in a temporal signal. Having similar features in both modalities could help to improve the multimodal representation by taking advantage of the correlation and other possible relations of both types of data.

3.2.1 Acceleration modality

First, from the accelerometer tri-axial data we take the raw data of each axis, the absolute values of each axis and the acceleration magnitude to obtain 7 different signals. Then these 7 signals are divided into 10-second time windows, from which statistical data such as mean, variance and power spectral density (PSD) are obtained, called *low-level features*. Subsequently, from the *low-level features*, the synchrony features are extracted as according the equations shown in Section 2.3. Following the same methodology used by Kapcak et al. [21], all features are correlated to the same axis, i.e, there is no cross-correlation between axes for any feature. The process to get all of synchrony features is shown in Figure 3.2, where in the last stage we can see the all types of features. So, for the 7 signals by axis we have 8 values for Correlation, Mutual information and time-lagged correlation and 4 values for Mimicry, that means that we have 196 feature values in total.

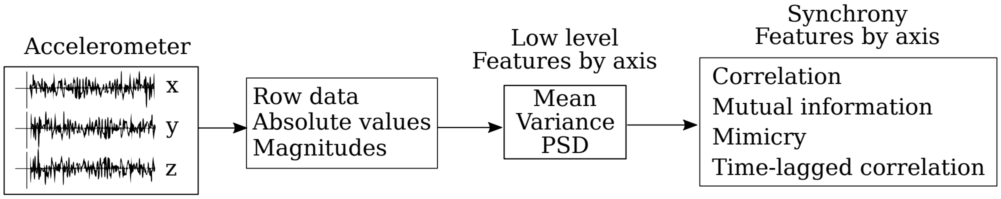


Figure 3.2: Process to extract synchrony features from acceleration modality.

3.2.2 Video modality

As a first attempt we will to reproduce the state of the art for the video modality. The approach uses features based on the position and location of the participants, these features are classified as follow.

1. The angle of location: which gives information on how one person is positioned with respect to the other.
2. The distance: which is calculated as the difference of the average Euclidean distance between the first n frames and the last n frames. This data is important because it is required to know how far or close a person who shows interest in the other person can be, both at the beginning of the appointment and at the end.
3. Movement distribution: this represents how often a person moves in all directions relative to the other person.

From these 3 classifications we derive the features that need to be measured and used as input to the classifier, the Table 3.2 summarizes these features.

Tabla 3.2: Video modality position-base features [36].

Features	**Description**
AVG-DIFANGLE	Average angle between the participant and the table
VARDIS	Distance variance
VAR-DIFANGLE	Variance of angle between the participant and the table
VARPOS	Position variance
VARPOS-OTHER	Variance in the position of the other person
DECRDIS	Decreasing in the distance
MOVDISTR	Movement distribution
MOVDISTR-OTHER	Distribution of movement of the other person
AVGDIS	Average distance

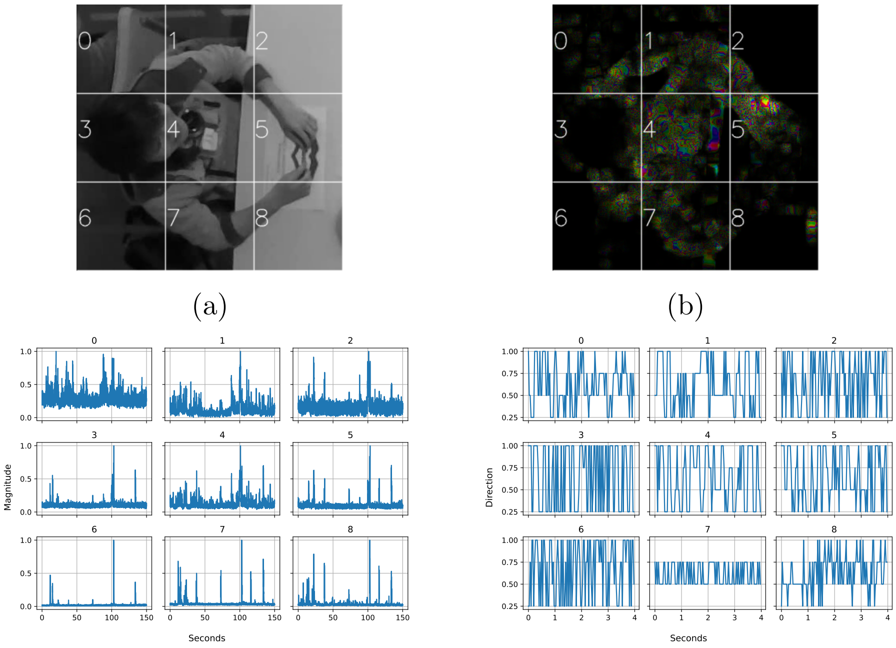
The above features proposed in the state of the art for this modality could not give us relevant information from our dataset. This was mainly due to the fact that these features are based on the distance variation and the relative position of the participants, which worked well for Veenstra et al. [36] standing scenario. However, in our dataset the participants were seated the whole time. Therefore, these features do not contain important information for our research. For this reason, we needed to also propose a new video baseline.

We propose to use the Optical Flow method to extract the video motion and then the motion synchrony features. The Optical Flow method is used in order to extract the motion contained between each frame on every speed date video, and then obtain the motion changes over time by concatenating the motion between the frames.

Two different types of signals are achieved from the optical flow, the **magnitude** and the **angle** of the pixel displacement vectors. The first gives us information about the movement intensity and the second movement direction. The average of the magnitudes is calculated and a histogram of the angles from which one direction of motion is chosen, giving a single intensity signal and a single direction signal. However, by averaging the magnitudes of all the pixels in the image and making a histogram of all the angles, we are losing local information in specific areas of the image.

In order not to lose local information, we will make a grid in the image to extract the optical flow from each cell, as shown in Figures 3.3a and 3.3b, where each cell roughly corresponds to a person's body part. For example, the hands will be contained mostly in the cells 2, 5 and 8, the head in 3 and 4, and the torso in cell number 4.

The optical flow average of each cell is calculated to obtain a single magnitude signal and a histogram to chose a single angle signal by cell. With this we will have the average data of intensity and direction of movement in each cell, which could give us an approximation of movement for specific body parts. This is shown in Figure 3.3c, the magnitude signal in



(a) (b)

(c) (d)

Figure 3.3: Image grid for local optical flow processing: (a) Grid on real frame, (b) Grid on magnitude of the optical flow, (c) Plots of magnitude by grid cell, and (d) Plots of direction by grid cell.

cells 0, 1 and 2 have more variation comparing with cells 6, 7 y 8, which indicates that the left hand, arm and shoulder move more than the right hand, arm and shoulder. Similarly, the angle signal represents the direction of movement, having a range of 360° possible directions. So we proceed to discretize these values in four directions with semantic meaning, i.e. Front (F), Back (B), Right (R) and Left (L) as shown in Figure 3.4. This is done for each frame of the video.

For this, a four bins histogram is created as follows:

1. bin 1, Front: angles between $[0, 45[$ and $[315, 360]$
2. bin 2, Right: angles between $[45, 135[$
3. bin 3, Behind: angles between $[135, 225[$
4. bin 4, Left: angles between $[225, 315[$

From this histogram, we choose the bin that contains the highest number of values by each grid cell and by 10 frames, so we will have a normalized single signal with four possible values. Figure 3.3d show us how the direction movement varies by each cell, that gives information on where each part of the body is moving towards. For example, in cell number 4 the direction varies from 0 to 1 with few intermediate values, which means that the head and torso are constantly moving back to front and with few movements to the right or left.

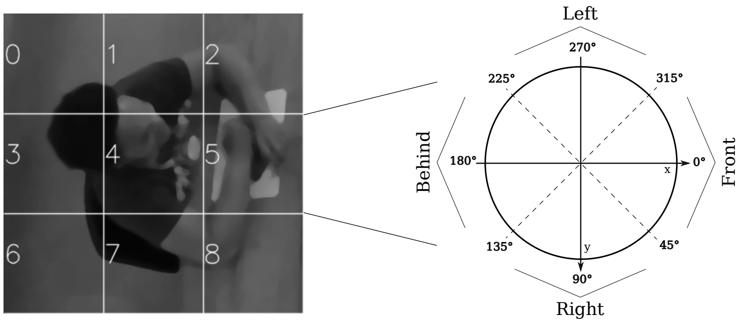


Figure 3.4: Discretizing angles in four directions with semantic meaning Front (F), Back (B), Right (R) and Left (L).

For this modality we also obtain the same synchrony features in Section 3.2. But in this case, we are correlating all optical flow signals from each grid cell against the other grid cells in the same participant video generating Individual features. We are not correlating against video data from the partner, i.e. we are no generating Mutual features. We hypothesize that the multimodal representation model can learn those Mutual features. So we have 4537 video synchrony features in total.

The whole process to extract the motion from optical flow video, is represented in the flow chart of Figure 3.5.

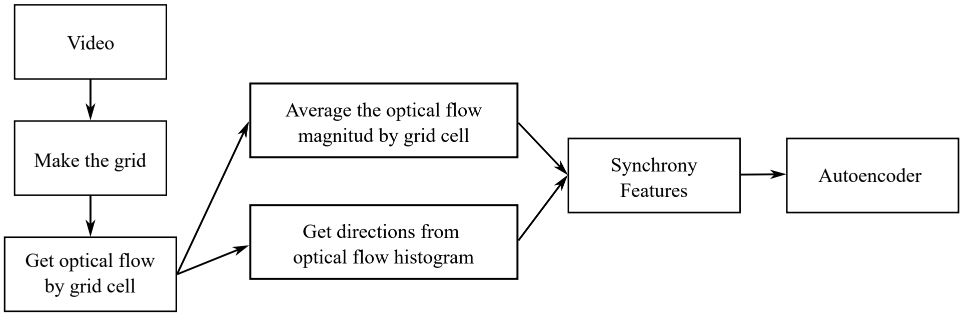

Video
Make the grid
Get optical flow by grid cell
Average the optical flow magnitud by grid cell
Get directions from optical flow histogram
Synchrony Features
Autoencoder

Figure 3.5: Motion extraction from video process.

3.3 Multimodal representation

The multimodal representation is done by implementing an autoencoder model. The idea is to take all the features from both modalities, concatenate them and use them as input to the decoding stage to compress them sufficiently to reconstruct the input data, and then take the latent space as *Multimodal Representation* for classification.

The Figure 3.6 shows the autoencoder architecture, which is is described as follows: We have two stages of layers: encoder and decoder, with exactly the same type of layers, same weights and same parameters. Both stages have one hidden layer and the encoder input layers is equal to the decoder output layer, and there is a layer that holds the latent

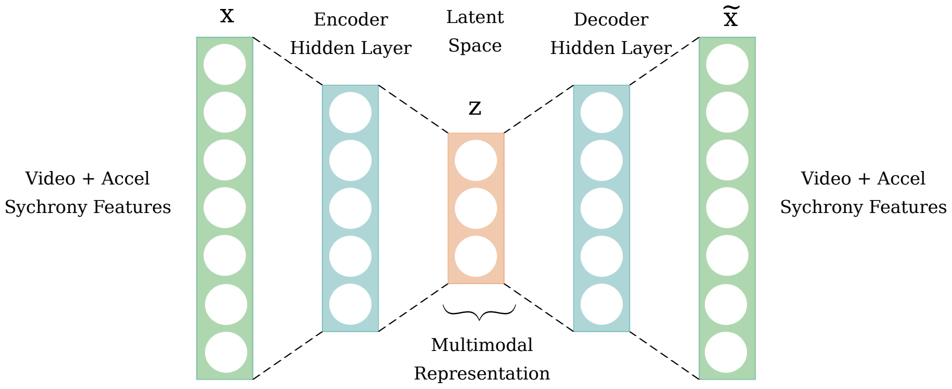


Figure 3.6: Autoencoder for proposed solution.

space. The hyperbolic tangent (tanh) was used as the activation function and MSE as loss function. Layer's dimension, droppout rate and learning rate were chosen by performing a Grid-Search parameter-tuning technique. For the latent space size we first used PCA keeping a 95% of the variance to have an approximation of the latent space size. Then, the parameter-tuner was given a range around this value. For the training, we take only the dates of day 1 and for validation the dates of day 2, and we will test it using day 3. The training dataset will contain both genders together.

The target in this case is for the autoencoder to learn a representation in latent space enough to reconstruct the input data. We will measure this reconstruction by evaluating the convergence of the mean square error (MSE).

3.4 Classification

The classification is implemented over the four types of individual interests and over the four types of mutual interest. What we want to observe is whether there is a difference between the types of interests that suggests that the nonverbal behavior in the multimodal representation is indicative of one of these types of attraction. We also want to see if the individual cases show differences with the mutual cases, i.e., to find evidence that mutual and individual interests are manifested by different nonverbal behaviors reflected in the classifier's performance. And finally, we want to understand whether separating the data between males and females has any implications for classification performance, i.e., to see how the difference between genders may suggest, in some way, that males and females express different nonverbal behaviors in relation to human attraction.

All questionnaire responses have been labeled using a 7-point Likert scale [20], that need to be binarized first to be used in classification. For this, we take each participant's scores for all participants dates to normalize them with the z-score normalization. Then, dates that have a positive value are labeled as *positive class* and those with negative values are labeled as *negative class*. Following this, mutual interactions are labeled as *positive class* when both individual dates have been labeled as *positive class*.

To reproduce the state of the art (unimodal baselines), first we will use all the features in the baselines to train the classifier for an individual modality. We will follow the same training methodology in the state of the art: using a 10-Fold cross-validation to have the AUC metric. We also want to understand if our video baseline performs better than the state-of-the-art baseline. For this, we will make a comparison between both baselines to know if our proposal is significantly better than the state of the art.

For the classification task in our proposal, we will take the latent space of the autoencoder trained with motion synchrony features to do the classification in order to compare with unimodal baselines. We propose to use 3-fold cross-validation instead of 10-fold cross-validation, because in the 10-fold approach there is data from the same person in the training and test set, and could cause an unrealistic prediction and may overfit the model. In 3-fold approach we will take two days for training and one day for validation, that means leave one day out. The classifier to be used in all classification experiments will be an *SVM*, the hyper-paramenters are tuned using a Grid-Seacrh method for each classification problem.

The cross-validation experiments will be done using data from both genders, from males only and from females only, as previous work showed that this separation affects the classification. Then, we collect the AUC metric for individual and mutual interests. We also implement a tuning parameter method in order to find the optimal model parameter values base on each classification problem and gender.

To evaluate whether the proposed solution provides a statistically significant improvement over the state of the art, we will perform this by collecting 30 values of AUC from different folds as Demšar [10] and Trawinski et al. [32] proposed, by gender and by problem classification. To do so, we will apply a 10x3-Fold cross-validation, i.e., to repeat the 3-Fold cross-validation 10 times. This is for baselines and our solution.

Having these AUC distributions for baselines and our solution, we will use a non-parametric hypothesis test called *Wilcoxon* to compare the mean of two distribution as was done by Rey et al. [30]. The null-hypothesis will be that the mean of AUC of the multimodal approach is less or equal than the mean of the state of the art. Therefore, the hypothesis we want to validate is that the mean of the multimodal approach is higher than the mean of the state of the art by rejecting the null-hypothesis.

Chapter 4

Results and analysis

4.1 Reproducing the state-of-the-art

4.1.1 Using baseline features

The results of the experiments performed, which are based on the state-of-the-art, are presented below. The *MatchNMingle (MnM)* dataset described in Section 2.2 was used in all cases.

First, we proceeded to replicate the experiments that gave the best results to Veenstra et al. [36] with the video modality. It is shown empirically that the variance in distance (VARDIS) has higher AUC values when the participants only want to exchange information.

In Table 4.1 the results for each classification problem are shown, using the same 10-fold cross-validation as presented by Veenstra et al. [36] but with MnM dataset. It can be observed that the highest AUC values are those related to problem *Friendly* in individual case and in *Match_SeeAgain/Match_Friendly* for mutual case, however, the highest value is 0.56 ± 0.12 which shows a low performance for these problems.

Another experiment with better performance for the same classification problem, which is stated in [36], is to use the position variance. It was suggested that the classifier can discriminate better between classes when using data only from males being themselves who vary their position (VARPOS). When using data only from women, it is their partners who vary their position (VARPOS-OTHER). That is, using only male data with VARPOS features, and using only female data with VARPOS-OTHER features, the classifier should show the best performance. The results of these experiments are shown in Table 4.2 and Table 4.3.

As shown in Tables 4.2 and 4.3, the performance of the classifiers does not vary much for each classification problem, and neither if they use data from males and females separately. In each case the performances are around random.

Tabla 4.1: Mean AUC±STD results with variance in distance with video modality using MnM dataset.

Problem	**Males ∪ Females**	**Males**	**Females**
SeeAgain	0.48 ± 0.07	0.54 ± 0.06	0.51 ± 0.06
Friendly	0.55 ± 0.08	0.49 ± 0.10	0.49 ± 0.13
Sexual	0.54 ± 0.10	0.46 ± 0.06	0.53 ± 0.10
Romantic	0.50 ± 0.06	0.51 ± 0.10	0.45 ± 0.11
Match_SeeAgain	0.47 ± 0.05	0.52 ± 0.08	0.56 ± 0.06
Match_Friendly	0.52 ± 0.05	0.56 ± 0.12	0.49 ± 0.08
Match_Sexual	0.53 ± 0.11	0.50 ± 0.11	0.52 ± 0.18
Match_Romantic	0.54 ± 0.06	0.49 ± 0.16	0.42 ± 0.12

Tabla 4.2: Mean AUC±STD results with position variance with video modality using MnM dataset.

Problem	**Males ∪ Females**	**Males**	**Females**
SeeAgain	0.49 ± 0.07	0.54 ± 0.09	0.58 ± 0.09
Friendly	0.49 ± 0.07	0.55 ± 0.09	0.58 ± 0.07
Sexual	0.56 ± 0.07	0.46 ± 0.11	0.44 ± 0.12
Romantic	0.53 ± 0.05	0.45 ± 0.09	0.49 ± 0.12
Match_SeeAgain	0.50 ± 0.09	0.51 ± 0.12	0.45 ± 0.13
Match_Friendly	0.48 ± 0.08	0.44 ± 0.12	0.47 ± 0.16
Match_Sexual	0.49 ± 0.08	0.48 ± 0.13	0.46 ± 0.15
Match_Romantic	0.46 ± 0.08	0.51 ± 0.16	0.50 ± 0.19

Tabla 4.3: Mean AUC±STD results with variance in the other person's position with the video modality using MnM dataset.

Problem	**Males ∪ Females**	**Males**	**Females**
SeeAgain	0.48 ± 0.06	0.45 ± 0.13	0.53 ± 0.10
Friendly	0.48 ± 0.07	0.47 ± 0.09	0.48 ± 0.11
Sexual	0.52 ± 0.07	0.53 ± 0.10	0.55 ± 0.11
Romantic	0.45 ± 0.05	0.50 ± 0.09	0.54 ± 0.11
Match_SeeAgain	0.48 ± 0.06	0.50 ± 0.15	0.58 ± 0.10
Match_Friendly	0.51 ± 0.08	0.54 ± 0.10	0.49 ± 0.08
Match_Sexual	0.44 ± 0.07	0.48 ± 0.13	0.42 ± 0.07
Match_Romantic	0.47 ± 0.12	0.43 ± 0.16	0.50 ± 0.13

The reason for under-performance in the above experiments is probably due to the nature of the data, and as explained before, in our dataset the participants were seated at all times, while for the scenario in [36] they were standing and can move at will around a high-table. Therefore, in the distance and position variation there is no relevant information that can be used to discriminate between classes, at least with this dataset. As a consequence, additionally to the original objectives of this these, we needed to propose an alternative approach for the video modality (see Section 3.2.2).

With the acceleration modality, we performed the same experiment by done Kapcak et al. [21], which consists of using all the features to train the classifier on each of the classification problems with 10-fold cross-validation, mixing males and females and separating them. Table 4.4 shows these results. It can be seen that the performance of the classifier is low in most cases, having AUC values around 0.5.

Tabla 4.4: Mean AUC±STD results with the acceleration mode for individual interest using MnM dataset.

Problem	**Males ∪ Females**	**Males**	**Females**
SeeAgain	0.53 ± 0.07	0.53 ± 0.11	0.50 ± 0.12
Friendly	0.51 ± 0.05	0.49 ± 0.07	0.50 ± 0.12
Sexual	0.52 ± 0.05	0.51 ± 0.09	0.59 ± 0.13
Romantic	0.47 ± 0.08	0.50 ± 0.09	0.45 ± 0.11

Also, the results were obtained for mutual interest, where the authors use only the data from both participants and do not make the separation between men and women. Table 4.5 shows these results. The classifier behavior in this case shows better performance than in the previous experiment.

Tabla 4.5: Mean AUC±STD results with the acceleration modality for mutual interest using MnM dataset.

Problem	**Males ∪ Females**
Match_SeeAgain	0.58 ± 0.06
Match_Friendly	0.59 ± 0.05
Match_Sexual	0.66 ± 0.07
Match_Romantic	0.60 ± 0.10

The results shown in the Tables 4.4 and 4.5 are similar to shown in the extended version of the baseline [35]. In our proposal we apply cross-validation based on 3-folding, taking one day out, so the AUC results are expected to decrease.

4.1.2 Using motion synchrony features only

Experiments were then performed using only the synchrony features in each modality. As in previous experiments, male and female data were used together and separately.

Table 4.6 shows the AUC results of the classifiers for each classification problem and using both genders. It can be seen that for video modality the performance behavior is similar to the previous case, the AUC values are over 0.5 but remain close to be random. The AUC values in the acceleration modality show that the classifier is, somehow, discriminating better in mutual problems, where AUC values are around 0.6.

Tabla 4.6: Mean AUC±STD results for both modalities with synchrony features with male and female data using MnM dataset.

Problem	Video	Acceleration
SeeAgain	0.56 ± 0.05	0.52 ± 0.05
Friendly	0.53 ± 0.06	0.52 ± 0.03
Sexual	0.54 ± 0.06	0.51 ± 0.06
Romantic	0.56 ± 0.06	0.50 ± 0.04
Match_SeeAgain	0.52 ± 0.03	0.59 ± 0.06
Match_Friendly	0.52 ± 0.10	0.60 ± 0.09
Match_Sexual	0.53 ± 0.09	0.64 ± 0.10
Match_Romantic	0.50 ± 0.09	0.64 ± 0.03

Tables 4.7 and 4.8 show a similar performance with the video modality, where only one gender is used, AUC values are close to be random in most of cases. Only Romantic interest is around 0.6 in both scenarios. For acceleration modality in both Tables 4.7 and 4.8, the highest AUC values are in Romantic and Sexual interests respectively.

Tabla 4.7: Mean AUC±STD results for both modalities with synchrony features and only male data using MnM dataset.

Problem	Video	Acceleration
SeeAgain	0.51 ± 0.11	0.53 ± 0.07
Friendly	0.47 ± 0.08	0.47 ± 0.08
Sexual	0.52 ± 0.08	0.51 ± 0.08
Romantic	0.59 ± 0.13	0.54 ± 0.09
Match_SeeAgain	0.50 ± 0.08	0.50 ± 0.08
Match_Friendly	0.48 ± 0.09	0.41 ± 0.12
Match_Sexual	0.51 ± 0.11	0.49 ± 0.13
Match_Romantic	0.44 ± 0.15	0.40 ± 0.13

4.1.3 Our proposed video baseline

To validate if our proposed video baseline outperforms the state-of-the-art, we classify using our new features. First we applied principal component analysis (PCA) on the features preserving 95% of the variance. Then, we classify using the same 10-fold cross

Tabla 4.8: Mean AUC±STD results for both modalities with synchrony features and only female data using MnM dataset.

Problem	**Video**	**Acceleration**
SeeAgain	0.58 ± 0.08	0.51 ± 0.09
Friendly	0.52 ± 0.16	0.54 ± 0.12
Sexual	0.53 ± 0.07	0.55 ± 0.11
Romantic	0.60 ± 0.10	0.49 ± 0.10
Match_SeeAgain	0.43 ± 0.12	0.50 ± 0.09
Match_Friendly	0.46 ± 0.09	0.53 ± 0.12
Match_Sexual	0.48 ± 0.12	0.49 ± 0.15
Match_Romantic	0.51 ± 0.16	0.43 ± 0.14

validation approach used in the state-of-the-art. For this test we are using data from both genders. Doing this, the Table 4.9 shows the comparison between state-of-the-art baseline and our proposed baseline.

Tabla 4.9: Mean AUC±STD results for comparison of state-of-the-art baseline (STD B) and our proposal (OUR B). ∗ indicates that the result significantly outperforms the video baseline with $p < 0.05$.

	Males U Females		**Males**		**Females**	
Target	**STA B**	**OUR B**	**STA B**	**OUR B**	**STA B**	**OUR B**
SeeAgain	0.46±0.065	0.50±0.075 ∗	0.54±0.057	0.43±0.094	0.44±0.040	0.58±0.077 ∗
Friendly	0.51±0.032	0.44±0.063	0.51±0.035	0.43±0.108	0.48±0.051	0.52±0.090
Sexual	0.44±0.041	0.52±0.084	0.57±0.060	0.46±0.057	0.56±0.040	0.51±0.115
Romantic	0.52±0.041	0.49±0.092	0.56±0.053	0.43±0.091	0.56±0.053	0.43±0.135
Match_SeeAgain	0.52±0.078	0.49±0.071	0.44±0.053	0.50±0.120	0.58±0.046	0.63±0.091
Match_Friendly	0.53±0.071	0.41±0.051	0.52±0.086	0.45±0.085	0.52±0.044	0.39±0.106
Match_Sexual	0.48±0.077	0.55±0.128	0.47±0.064	0.59±0.138 ∗	0.53±0.072	0.58±0.139
Match_Romantic	0.55±0.099	0.66±0.045 ∗	0.48±0.053	0.53±0.104	0.50±0.096	0.46±0.147

In Table 4.9 we can see that mean AUC values comparing both baselines video approaches, where only four cases are outperformed statistically. There are cases where there is no significant improvement, but the mean AUC is higher by 18% in the case of the *Sexual* interest using both genders, also for the case of *Match_SeeAgain* using female data, the mean AUC is higher by 8%. This percentage increase in the mean AUC corresponds to a high standard deviation, therefore no statistically significant improvement is evident. Although we have a significant improvement for only four cases, we keep using this method as baseline, to understand whether merging this baseline with the acceleration modality will result in a significant improvement.

4.2 Multimodal representation

A first attempt to use a multimodal approach, we used a *Majority Voting* classifier to understand the implications of joining multi-modalities in a simple manner. So, the Table 4.10 shows the results of this approach, were we can see that for some types of interest and genders, the classifier can perform over random, having the high AUC value in SeeAgain using both genders, Match_Romantic using males and See_Again using females. It should be noted that this classifier depends on the previous classifiers, so if the previous classifier behave randomly Majority Voting will behaves randomly as well. For this reason, in our solution we propose an early fusion method that can take advantage of features.

Tabla 4.10: Mean AUC±STD results *Majority Voting* classifier with synchrony features using MnM dataset.

Problem	Males U Females	Males	Females
SeeAgain	0.56 ± 0.06	0.50 ± 0.11	0.59 ± 0.07
Friendly	0.50 ± 0.05	0.53 ± 0.06	0.51 ± 0.16
Sexual	0.47 ± 0.07	0.51 ± 0.09	0.47 ± 0.07
Romantic	0.44 ± 0.07	0.42 ± 0.14	0.41 ± 0.10
Match_SeeAgain	0.46 ± 0.05	0.48 ± 0.10	0.56 ± 0.12
Match_Friendly	0.46 ± 0.11	0.54 ± 0.08	0.52 ± 0.11
Match_Sexual	0.44 ± 0.10	0.48 ± 0.11	0.52 ± 0.12
Match_Romantic	0.47 ± 0.07	0.60 ± 0.13	0.49 ± 0.17

As described in Section 1.3, our proposal is to use an autoencoder as early fusion method to generate an multimodal representation, trained with motion synchrony features from MnM dataset. Figure 4.1 depicts the trending of error function for training and for validation. We see that the loss function tends to converge in both cases, however the validation loss diverges after 1862 epochs, so we decided to make an early stop and take the model trained up to this epoch. With this we are also reducing the overfitting in the model because the training loss tend to converge even more.

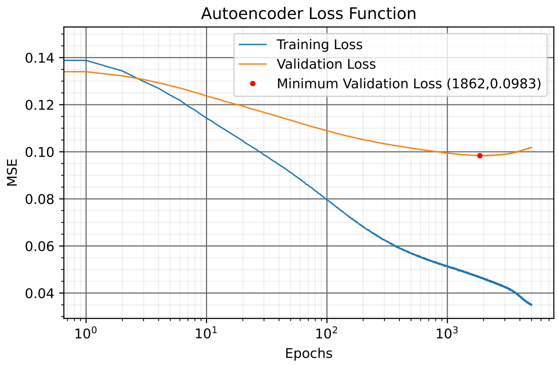


Figure 4.1: Autoencoder Training and validation Loss functions.

From Figure 4.1 it can be observed that the reconstruction error (validation losses) does tend to converge, however, the error could be expected to decrease even more, it seems that the autoencoder is not being able to completely reconstruct the input, the value of MSE can get even closer to zero.

According to Figure 4.2, using the testing dataset, we have the minimal MSE value when we use both genders, indicating that the model is actually learning a representation having the better performance using both genders. This suggests that the pattern of behavior the model is learning exists when both genders are together. The MSE decreases using only one gender. This indicates that behavioral patterns for each gender could be different and the multimodal representation could change separating by genders. So, the autoencoder is achieving some representation that is able to reconstruct the input, but is not enough to achieve a clear division between classes.

The autoencoder was trained to reconstruct an input without any label information, so the learning of the model has been based only on the motion synchrony features. The reason for doing so is to have a pure representation of the data without the influence of the labels, which could force a representation that does not fit the problem as such.

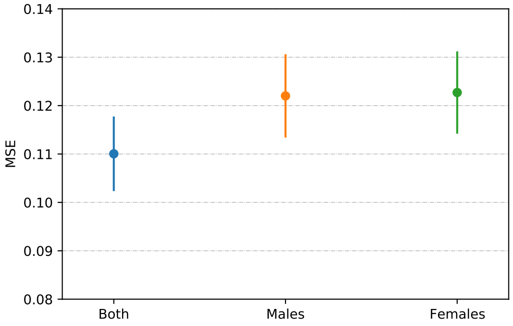
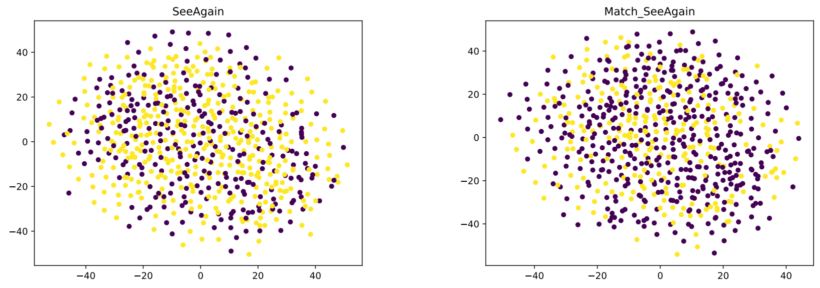


Figure 4.2: Mean MSE and Standard Deviation for Autoencoder using testing dataset.

To analyze the separability in the latent space, we use the *t-Distributed Stochastic Neighbor Embedding (t-SNE)* over the latent space for the main problem (SeeAgain). This is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets [34].

Figure 4.3 shows that there is no clear separation of classes in the latent space when it is projected onto 2D dimensional space. However, we cannot conclude that there is no any separability at all, because we are reducing high dimensional space (500 dimensions) to only 2 dimension, so we could are missing important information in other dimensions. What Figure 4.3 may be telling us is that labels are not accurately representing the attraction problem as it is, possibly more modalities are needed or more non-verbal behavioral features may need to be added. If this is correct, this missing separability will be shown in classification experiments.



(a) (b)

Figure 4.3: t-SNE visualizing classes in Autoencoder latent space. (a) SeeAgain problem and (b) Match_SeeAgain problem.

4.3 Results of classification

The results of the classification obtained for the two baselines and for our proposed multimodal approach are shown below. These were done using a 10x3-fold cross-validation. As we described in Section 3.4 we apply the Wilcoxon test to verify whether our proposed solution significantly outperforms the baselines. The Table 4.11 shows the mean AUC by classification problem for each modality approach using data from both genders.

Tabla 4.11: Mean AUC±STD for unimodal and multimodal approaches using *both genders*, synchrony features and MnM dataset. $*$ indicates that the result significantly outperforms the video baseline with $p < 0.05$, $\circledast$ indicates that the result significantly outperforms the acceleration baseline with $p < 0.05$ and $\dagger$ indicates that the result significantly outperforms both baselines with $p < 0.05$.

Problem	**Video**	**Acceleration**	**Multimodal**
SeeAgain	0.50±0.021	0.48±0.052	0.52±0.032 $\circledast$
Friendly	0.47±0.020	0.50±0.004	0.48±0.019 $*$
Sexual	0.42±0.006	0.50±0.000	0.49±0.037 $*$
Romantic	0.49±0.044	0.50±0.000	0.52±0.040 $\dagger$
Match_SeeAgain	0.49±0.049	0.48±0.033	0.49±0.020
Match_Friendly	0.47±0.050	0.52±0.029	0.54±0.051 $\dagger$
Match_Sexual	0.56±0.056	0.58±0.032	0.47±0.060
Match_Romantic	0.52±0.114	0.49±0.068	0.54±0.072 $\circledast$

For individual interests the multimodal approach shows that it's able to outperform the four cases, in particular, the case of Romantic is outperforming both baselines. For mutual interests, the multimodal approach is outperforming Match_Friendly interest only.

Using both genders our hypothesis is statistically validated over both baselines only for

Romantic individual interest and for Match_Friendly mutual interest, with a performance over random. Also, for SeeAgain and Match_Romantic interest, our hypothesis is only validated over the acceleration baseline. Also, for Romantic and Match_Friendly, both modalities are correlated and the behavior captured in both can mutually complement each other. This suggests that for that particular type of interest both men and women could have similarities in non-verbal behavior, i.e., their behavior looks similar when they want to see someone again and when they have a friendly or romantic interest. The cases where the multimodal approach outperforms the video baseline have the problem that the mean AUC is around random. Therefore, even when our hypothesis can be demonstrated, all classifiers, including those of the baselines do not have enough or relevant information to learn the difference within attraction patterns.

We also did the same experiment above using data from males only, this in order to understand if there any differences from the male's perspective only, as stated in the state-of-the-art. Table 4.12 shows the mean AUC results for this case. We see that the multimodal approach only outperforms the video modality in one individual case (Romantic) and two cases of mutual interests. The acceleration modality is outperformed only in Match_Romantic interest.

Tabla 4.12: Mean AUC±STD for unimodal and multimodal approaches using *males* only, synchrony features and MnM dataset. $*$ indicates that the result significantly outperforms the video baseline with $p < 0.05$, $\circledast$ indicates that the result significantly outperforms the acceleration baseline with $p < 0.05$ and † indicates that the result significantly outperforms both baselines with $p < 0.05$.

Problem	**Video**	**Acceleration**	**Multimodal**
SeeAgain	0.50±0.000	0.53±0.062	0.50±0.043
Friendly	0.45±0.019	0.49±0.047	0.44±0.023
Sexual	0.46±0.072	0.51±0.040	0.50±0.035
Romantic	0.48±0.008	0.50±0.064	0.53±0.064 $*$
Match_SeeAgain	0.41±0.041	0.53±0.082	0.48±0.054 $*$
Match_Friendly	0.44±0.039	0.49±0.025	0.51±0.070 $*$
Match_Sexual	0.51±0.073	0.50±0.000	0.51±0.108
Match_Romantic	0.56±0.098	0.49±0.014	0.51±0.009 $\circledast$

What we can see here is that for males the video baseline has random performance in most of cases. Only for Math_Romantic it's showing a performance over random, while the acceleration baseline has an over-random performance for SeeAgain, Sexual and Match_SeeAgain interest. What this suggests is that, for males, the motion capture by accelerometers is more important than motion capture by video. This may mean that whole body or trunk movements are more relevant in males.

There is no case where our hypothesis is demonstrated over both baselines. There are only two cases where the hypothesis is validated, those are Romantic and for Match_Friendly over video baseline and Match_Romantic over acceleration baseline.

Using females data only the results are also difference. Figure 4.13 shows the mean AUC for all classification problems. Where we can see that for individual cases the multimodal approach outperforms both baselines for Sexual interest only. In mutual interest problems the multimodal approach outperforms both baselines in two cases, Match_SeeAgain and Match_Friendly. For other two mutual interests only the acceleration baseline is outperformed.

Tabla 4.13: Mean AUC±STD for unimodal and multimodal approaches using *females* only, synchrony features and MnM dataset. $*$ indicates that the result significantly outperforms the video baseline with $p < 0.05$, $\circledast$ indicates that the result significantly outperforms the acceleration baseline with $p < 0.05$ and † indicates that the result significantly outperforms both baselines with $p < 0.05$.

Problem	**Video**	**Acceleration**	**Multimodal**
SeeAgain	0.50±0.021	0.48±0.079	0.49±0.054
Friendly	0.46±0.057	0.56±0.069	0.46±0.036
Sexual	0.48±0.049	0.47±0.073	0.52±0.066 †
Romantic	0.51±0.011	0.50±0.034	0.47±0.065
Match_SeeAgain	0.39±0.044	0.46±0.050	0.52±0.028 †
Match_Friendly	0.44±0.087	0.48±0.031	0.54±0.076 †
Match_Sexual	0.55±0.068	0.45±0.058	0.53±0.071 $\circledast$
Match_Romantic	0.45±0.035	0.45±0.039	0.50±0.107 $\circledast$

These results suggest that for females both modalities are important for Sexual, Match_SeeAgain and Match_Friendly interests. The multimodal approach is showing that both modalities are correlated. For this reason the hypothesis is demonstrated with a performance over random. Also, for Match_Sexual and Match_Romantic interest the hypothesis is only demonstrated over acceleration, which suggests, unlike males, that for females motion captured by video it's more important than acceleration. Movement captured by video as hands movements could be decisive for females attraction.

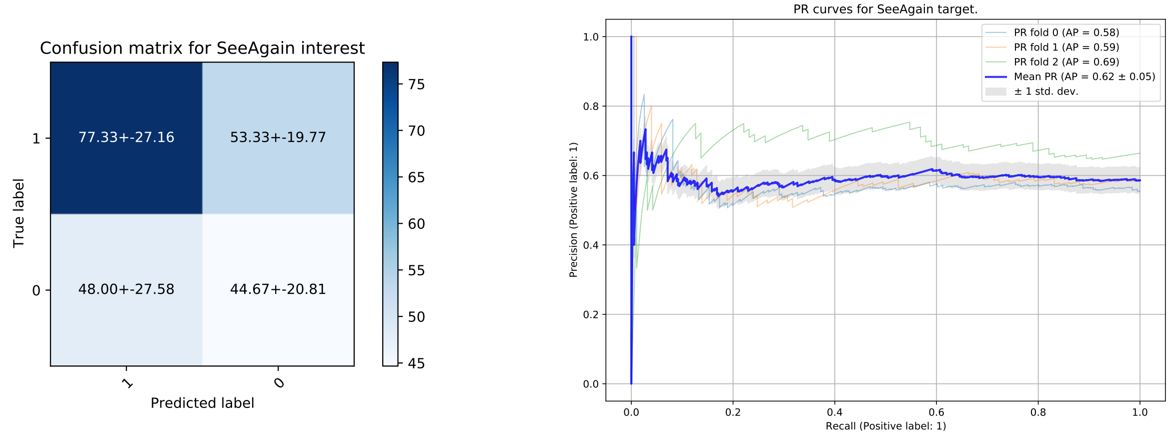
4.4 Further analysis

Analyzing the main problem (SeeAgain) for the individual case, we can see that the multimodal approach outperforms the video baselines when we use data from both genders. Even though there is an improvement in performance, the mean AUC is around 0.52 which means that the classifier is almost random. If we take at look at the confusion matrix and the PR curve in Figure 4.4, we see that the TP in confusion matrix are higher than

FP and FN, which means that the classifier is trying to separate positive class. However, when we look at the PR curve, the precision remains nearly constant even when recall is low or high, this means that there is no threshold where both precision and recall are close to one. The behavior of the PR curve shows that although there is some learning by the classifier, it is not able to find the optimal threshold for class separability.

For Match_SeeAgain, the multimodal approach only outperforms both baselines using females only and the acceleration baseline using males only. It is important to highlight that mutual cases have this out-performance using the data from only one participant. Using females the mean AUC of multimodal approach is around 0.52, while AUC values of baselines are less than 0.5.

Furthermore, we also evaluate SeeAgain for the mutual interest, instead of the independent one. The PR curve of Figure 4.5b shows a low precision and recall, because the classifier predicts mostly positive classes. This is shown in confusion matrix in Figure 4.5a where $TP + FP$ has the highest number. For this case and the other mutual cases, we need to consider the unbalance between classes with a difference of about 36:64 (%), where the positive class has the least number and the classifier weights were adjusted according to this unbalance. This unbalance is causing false positive predictions to occur more often than true ones.

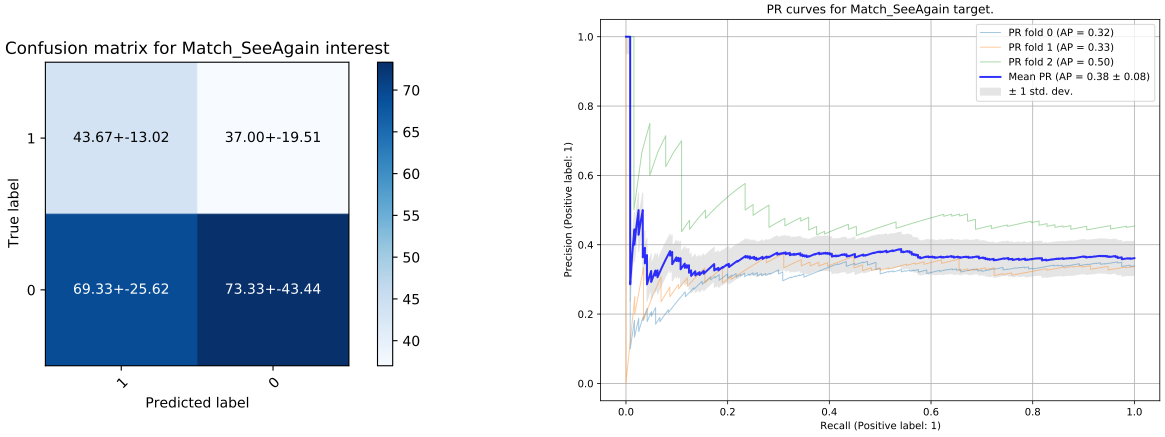


(a)

(b)

Figure 4.4: (a) Average confusion matrix and (b) Precision-recall curve for SeeAgain individual interest using both genders. Positive to Negative class balance percentage is 58:42 (%). Positive class is 1.

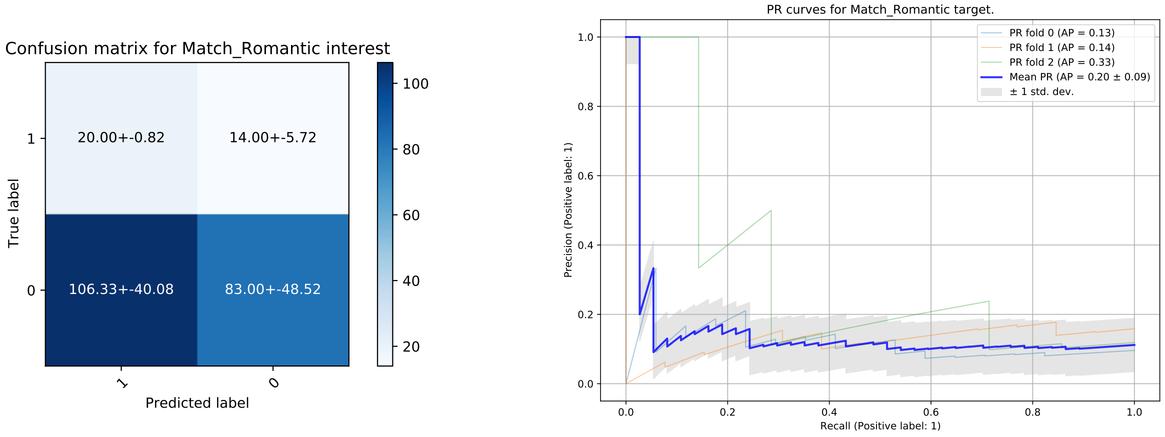
Another problem to analyze is the Romantic interest, because in all cases one or both baselines are outperformed with a mean AUC greater than random. The highest AUC occurs when both genders are present in the mutual case, AUC values is around 0.54. In Figure 4.6a we can see in the confusion matrix a similar behavior than Match_SeeAgain, where the high FP are causing a low precision even when the recall is low or high, this can be seen in the PR curve in Figure 4.6b.



(a)

(b)

Figure 4.5: (a) Average confusion matrix and (b) Precision-recall curve for SeeAgain mutual interest using both genders. Positive to Negative class balance percentage is 36:64 (%). Positive class is 1.



(a)

(b)

Figure 4.6: (a) Average confusion matrix and (b) Precision-recall curve for Romantic mutual interest using both genders. Positive to Negative class balance percentage is 15:85 (%). Positive class is 1.

If we take a look at individual romantic interests using only males, the performance improves with respect to the video baseline. Here we can see that precision tends to increase when recall also increases according to Figure 4.7. This result is important because it means that both FP and FN tend to go down. So, in this case the classifier is trying to separate both classes even when the dataset is unbalanced.

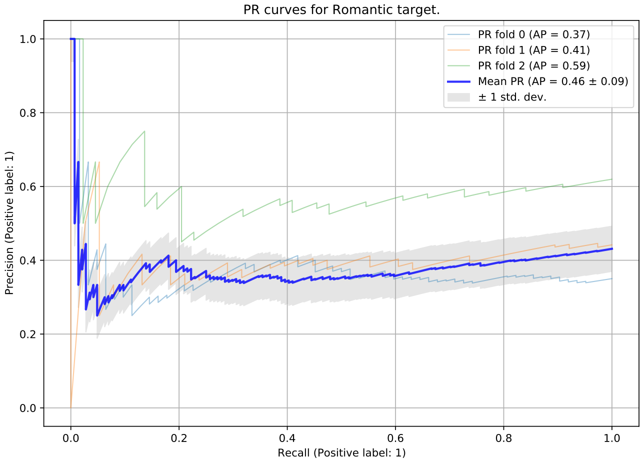
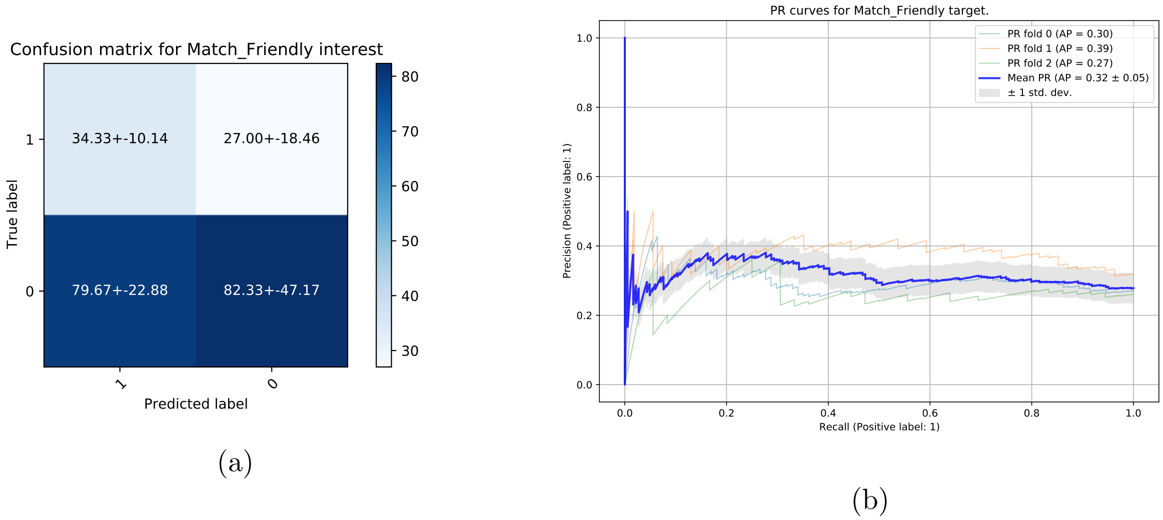


Figure 4.7: Precision-recall curve for Romantic individual interest using using males only. Positive to Negative class balance percentage is 43:57 (%). Positive class is 1.

The multimodal approach also outperforms both baselines for Mutual Friendly interests, using both genders and females only with an AUC around 0.54. However, although the performance is statistically better, for the case of females only, the improvement is over random classifiers. Whereas if both genders are used, there is a significant improvement over the baseline video whose AUC is greater than 0.5. For this particular case,the confusion matrix of Figure 4.8a exhibits a similar behavior than the previous cases, where the classifier is trying to find a threshold for class separability. This is shown in the PR curve of Figure 4.8b where at a recall value around 0.2 the precision trends to increase, but then decreases having a low AP value as indicator of low performance for the positive class.



(a)

(b)

Figure 4.8: (a) Average confusion matrix and (b) Precision-recall curve for Friendly mutual interest using using males only. Positive to Negative class balance percentage is 27:73 (%). Positive class is 1.

Chapter 5

Conclusions and future work

5.1 Conclusions

1. The state of the art video baseline proposes to use features based on the position and relative location of the participants, however the results presented in the replication of this baseline show that the classifiers behave in a random way, because these features cannot be obtained from our dataset.

2. A new video baseline based on synchrony features has been proposed and implemented, which had a better performance than the previous baseline for some particular cases, bringing the performance of the classifier closer to an AUC around 0.6, which shows that the behavior of the classifier is no longer random.

3. By performing experiments with the unimodal baselines using only motion synchrony features, the classifiers showed an increase in the average performance and did not behave randomly.

4. The results presented show that there is a statistically significant improvement of the multimodal approach against the baselines. In some cases it improves both baselines while in others it improves only one of them.

5. It was shown that the autoencoder does learn a representation capable of reconstructing the input containing the motion synchrony features. The latent space where the multimodal features are represented shows that the separability between the classes defined in the dataset labels is not observable in two dimensions.

6. The individual classification problems showed better performance in the presence of both genders, both SeaAgain and Romantic had higher average performances than random and showed a significant improvement over baselines. This indicates that when a man or a woman wants to see a person again or has a romantic interest, the nonverbal behavior shows similarities between both genders.

7. Using males information only the classifiers were not able, in most cases, to outperform the acceleration baseline. This could be an indicative that movement taken by wearable sensors, such as entire body movements, are most important in males than in females.

8. The results with data from women only show that the classifiers can outperform both the video and acceleration baselines. That means that both modalities contain important non-verbal information about attraction. This could indicate that both modalities are needed for females, the movements captured on video, such as hands movements, could be correlated to the motion capture by the wearable sensors. But this depends on what type of interest is under study.

9. According to the results both males and females show different classifier performances, which put in evidence that both genders have different non-verbal behaviors. New research should pursuit this premise.

10. Every type of attraction interest shows different performance results in all cases. This suggest that non-verbal behaviors are different depending on the interest type.

11. The classifiers show that in some way there is class separability with performance over but closely to be random.

5.2 Future work

1. Do a deeper analysis of the attraction problem to understand if the label present in our dataset can be more accurate based on other approaches and psychological research.

2. Training the autoenconder in a guided way. That means the model learns a representation based on the label.

3. Include more modalities that can capture other types of behaviors that may result in more reliable information about human attraction.

4. A method can be proposed to capture a particular body movement to understand whether it is a cue of human attraction.

5. Methodology to improve the classification based on another Machine Learning approaches should be proposed.

6. A new method that learns human attraction features based from raw signals based on attention methods as Transformers should be proposed.

Bibliography

[1] P.K. Atrey, M.A. Hossain, A. El Saddik, et al M.S. Kankanhalli. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16(6):345–379, 2010. doi: 10.1007/s00530-010-0182-0. iii, 13, 14

[2] Pierre Baldi. Autoencoders, Unsupervised Learning, and Deep Architectures. *ICML Unsupervised and Transfer Learning*, pages 37–50, 2012. ISSN 0899-7667. doi: 10.1561/2200000006. 16

[3] T. Baltrusaitis, C. Ahuja, et al L.-P. Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. doi: 10.1109/TPAMI.2018.2798607. iii, 12, 14, 15

[4] Yoshua Bengio, Aaron Courville, et al Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. ISSN 01628828. doi: 10.1109/TPAMI.2013.50. 14

[5] L. Cabrera-Quiros, E. Gedik, et al H. Hung. Estimating self-assessed personality from body movements and proximity in crowded mingling scenarios. pages 238–242, 2016. doi: 10.1145/2993148.2993170. URL `https://www.scopus.com/inward/record.uri?eid=2-s2.0-85016611614&doi=10.1145%2f2993148.2993170&partnerID=40&md5=de48e30c2167e0ace85f619b688c290b`. cited By 3. 11

[6] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L.V.D. Meij, et al H. Hung. The MatchNMingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing*, 2018. doi: 10.1109/TAFFC.2018.2848914. 8, 9

[7] Mohamed Chetouani, Emilie Delaherche, Guillaume Dumas, et al David Cohen. Interpersonal synchrony: From social perception to social interaction. *Social Signal Processing*, (1988):202–212, 2017. doi: 10.1017/9781316676202.015. 10, 11, 12

[8] Felipe L.A. Conceiç ao, Flávio L.C. Pádua, Anisio Lacerda, Adriano C. Machado, et al Daniel H. Dalip. Multimodal data fusion framework based on autoencoders for top-N recommender systems. *Applied Intelligence*, 49(9):3267–3282, 2019. ISSN 15737497. doi: 10.1007/s10489-019-01430-7. 17

[9] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, et al David Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, 2012. ISSN 19493045. doi: 10.1109/T-AFFC.2012.12. iii, 11, 12

[10] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006. ISSN 15337928. 28

[11] S.D. Farley. Nonverbal reactions to an attractive stranger: The role of mimicry in communicating preferred social distance. *Journal of Nonverbal Behavior*, 38(2):195–208, 2014. doi: 10.1007/s10919-014-0174-4. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-84898538587&doi=10.1007%2fs10919-014-0174-4&partnerID=40&md5=e5b215430cd5dbd13f052069e5dd01c0. cited By 21. 3

[12] Ekin Gedik et al Hayley Hung. Detecting conversing groups using social dynamics from wearable acceleration: Group size awareness. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(4), December 2018. doi: 10.1145/3287041. URL https://doi.org/10.1145/3287041. 11

[13] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, et al Stefan Scherer. Importance-based multimodal autoencoder, 2021. URL https://openreview.net/forum?id=4jXnFYaDOuD. 17

[14] N. Guéguen. Mimicry and seduction: An evaluation in a courtship context. *Social Influence*, 4(4):249–255, 2009. doi: 10.1080/15534510802628173. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-75149131561&doi=10.1080%2f15534510802628173&partnerID=40&md5=f62c9a9aae9b0b093b3f9a9e57c020ca. cited By 35. 3

[15] Dan Hu, Han Zhang, Zhengwang Wu, Fan Wang, Li Wang, J. Keith Smith, Weili Lin, Gang Li, et al Dinggang Shen. Disentangled-Multimodal Adversarial Autoencoder: Application to Infant Age Prediction with Incomplete Multimodal Neuroimages. *IEEE Transactions on Medical Imaging*, 39(12):4137–4149, 2020. ISSN 1558254X. doi: 10.1109/TMI.2020.3013825. 17

[16] Jin Huang et al C.X. Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, 2005. doi: 10.1109/TKDE.2005.50. iii, 18

[17] A Cecile JW Janssens et al Forike K Martens. Reflection on modern methods: Revisiting the area under the roc curve. *International journal of epidemiology*, 49 (4):1397–1403, 2020. 18

[18] Natasha Jaques, Sara Taylor, Akane Sano, et al Rosalind Picard. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. *2017 7th International Conference on Affective*

Computing and Intelligent Interaction, ACII 2017, 2018-Janua:202–208, 2018. doi: 10.1109/ACII.2017.8273601. 16

[19] Miao Jiaju et al Zhu Wei. Precision–recall curve (prc) classification trees. https://link.springer.com/article/10.1007/s12065-021-00565-2#citeas, April 2021. 19

[20] Ankur Joshi, Saket Kale, Satish Chandel, et al D Kumar Pal. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396, 2015. 27

[21] O. Kapcak, J. Vargas-Quiros, et al H. Hung. Estimating Romantic, Social, and Sexual Attraction by Quantifying Bodily Coordination using Wearable Sensors. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2019*, pages 154–160, 2019. ISBN 9781728138916. doi: 10.1109/ACIIW.2019.8925137. 3, 7, 8, 12, 23, 31

[22] J.C. Karremans et al T. Verwijmeren. Mimicking attractive opposite-sex others: The role of romantic relationship status. *Personality and Social Psychology Bulletin*, 34 (7):939–950, 2008. doi: 10.1177/0146167208316693. 3

[23] Walter Hugo Lopez Pinaya, Sandra Vieira, Rafael Garcia-Dias, et al Andrea Mechelli. Autoencoders. *Machine Learning: Methods and Applications to Brain Disorders*, pages 193–208, 2019. doi: 10.1016/B978-0-12-815739-8.00011-0. iii, 15, 16, 17

[24] Anmol Madan, Ron Caneel, et al Alex Pentland. Voices of attraction, 2004. 7

[25] Jan Michalsky et al Heike Schoormann. Pitch convergence as an effect of perceived attractiveness and likability, 2017. 7

[26] M.C. Nanninga, Y. Zhang, N. Lehmann-Willenbrock, Z. Szlávik, et al H. Hung. Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry. volume 2017-January, pages 206–215, 2017. doi: 10.1145/3136755.3136811. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85045474277&doi=10.1145%2f3136755.3136811&partnerID=40&md5=2a4fdc0d23ff8d0734c08f31b1fdb4f3. cited By 10. 3

[27] Elad Plaut. From principal subspaces to principal components with linear autoencoders, 2018. 16

[28] F. Ramseyer et al W. Tschacher. Nonverbal synchrony in psychotherapy: Coordinated body movement reflects relationship quality and outcome. *Journal of Consulting and Clinical Psychology*, 79(3):284–295, 2011. doi: 10.1037/a0023419. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-79958845401&doi=10.1037%2fa0023419&partnerID=40&md5=5429a7d2207b2374d6d454cb7e794ade. cited By 232. 3, 11

[29] R. Ranganath, D. Jurafsky, et al D. McFarland. It's not you, it's me: Detecting flirting and its misperception in speed-dates. In *EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009*, pages 334–342, 2009. 3, 7

[30] Denise Rey et al Markus Neuhäuser. *Wilcoxon-Signed-Rank Test*, pages 1658–1659. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-04898-2. doi: 10.1007/978-3-642-04898-2_616. URL https://doi.org/10.1007/978-3-642-04898-2{_}616. 28

[31] teletica.com. ¿debe el mercado de los sitios de citas preocuparse por facebook? — teletica. https://www.teletica.com/internacional/debe-el-mercado-de-los-sitios-de-citas-preocuparse-por-facebook_249211. (Accessed on 12/01/2020). 1

[32] Bogdan Trawinski, Magdalena Smetek, Zbigniew Telec, et al Tadeusz Lasota. Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. *International Journal of Applied Mathematics and Computer Science*, 22 (4):867–881, 2012. ISSN 1641876X. doi: 10.2478/v10006-012-0064-z. 28

[33] W. Tschacher, G.M. Rees, et al F. Ramseyer. Nonverbal synchrony and affect in dyadic interactions. *Frontiers in Psychology*, 5(NOV), 2014. doi: 10.3389/fpsyg.2014.01323. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-84923620534&doi=10.3389%2ffpsyg.2014.01323&partnerID=40&md5=44f9b8202d14188f270c233b3a31e69d. cited By 65. 11

[34] Laurens van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15(93):3221–3245, 2014. URL http://jmlr.org/papers/v15/vandermaaten14a.html. 35

[35] Jose David Vargas Quiros, Oyku Kapcak, Hayley Hung, et al Laura Cabrera-Quiros. Individual and joint body movement assessed by wearable sensing as a predictor of attraction in speed dates. *IEEE Transactions on Affective Computing*, pages 1–1, 2021. doi: 10.1109/TAFFC.2021.3138349. 8, 31

[36] A. Veenstra et al H. Hung. Do they like me? Using video cues to predict desires during speed-dates. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 838–845, 2011. ISBN 9781467300629. doi: 10.1109/ICCVW.2011.6130339. v, 1, 3, 8, 12, 24, 29, 30

[37] A. Vijayalakshmi et al P. Mohanaiah. *Literature Survey on Emotion Recognition for Social Signal Processing*, volume 614. 2020. ISBN 9789811506253. doi: 10.1007/978-981-15-0626-0_29. 1

[38] A. Vinciarelli. *Introduction: Social signal processing.* 2017. ISBN 9781316676202. doi: 10.1017/9781316676202.001. 1

[39] Alessandro Vinciarelli, Maja Pantic, et al Herv Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743 – 1759, 2009. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2008.11.007. URL http://www.sciencedirect.com/science/article/pii/S0262885608002485. 1