FACIAL ANALYSIS OF DYADIC INTERACTIONS USING MULTIPLE INSTANCE LEARNING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

By
Dersu Giritlioğlu
September 2021

Facial Analysis of Dyadic Interactions Using Multiple Instance Learning
By Dersu Giritlioğlu
September 2021

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Hamdi Dibeklioğlu(Advisor)

Shervin Rahimzadeh Arashloo

Albert Ali Salah

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

FACIAL ANALYSIS OF DYADIC INTERACTIONS USING MULTIPLE INSTANCE LEARNING

Dersu Giritlioğlu
M.S. in Computer Engineering
Advisor: Hamdi Dibeklioğlu
September 2021

Interpretation of nonverbal behavior is vital for a reliable analysis of social interactions. To this end, we automatically analyze facial expressions of romantic couples during their dyadic interactions, for the first time in the literature. We use a recently collected romantic relationship dataset, including videos of 167 couples while talking on a conflicting case and a positive experience they share. To distinguish between interactions during positive experience and conflicting discussions, we model facial expressions employing a deep multiple instance learning (MIL) framework, adapted from the anomaly detection literature. Spatio-temporal representation of facial behavior is obtained from short video segments through a 3D residual network and used as the instances in MIL bag formations. The goal is to detect conflicting sessions by revealing distinctive facial cues that are displayed in short periods. To this end, instance representations of positive experience and conflict sessions are further optimized, so as to be more separable using deep metric learning. In addition, for a more reliable analysis of dyadic interaction, facial expressions of both subjects in the interaction are analyzed in a joint manner.

Our experiments show that the proposed approach reaches an accuracy of 71%. In addition to providing comparisons to several baseline models, we have also conducted a human evaluation study for the same task, employing 6 participants. The proposed approach performs 5% more accurately than humans as well as outperforming all baseline models. As suggested by the experimental results, reliable modeling of facial behavior can greatly contribute to the analysis of dyadic interactions, yielding a better performance than that of humans.

Keywords: Dyadic Interaction, Behavior Analysis, Facial Expression, Multiple Instance Learning, Metric Learning, Deep Learning.

ÖZET

İKİLİ ETKİLEŞİMLERDE ÇOKLU ÖRNEKLE ÖĞRENME KULLANILARAK YÜZ İNCELEMESİ

Dersu Giritlioğlu
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Hamdi Dibeklioğlu
Eylül 2021

Sözsüz davranışların yorumlanması, sosyal etkileşimlerin güvenilir şekilde incelenmesinde büyük önem taşımaktadır. Bu amaçla, literatürde ilk kez, romantik çiftlerin ikili etkileşimlerinde yüz ifadelerini otomatik olarak incelemekteyiz. Çalışmamızda, 167 çiftin bir anlaşmazlıklarını ve paylaştıkları olumlu bir deneyimi konuştukları videoları içeren, yakın zamanda toplanmış romantik ilişki veri kümesi kullanılmaktadır. Olumlu deneyim ve anlaşmazlıklar sırasındaki etkileşimleri ayırt etmek için, anomali tespiti literatüründen uyarlanmış, Derin Çoklu Örnekle Öğrenme (MIL) çatısı kullanılarak yüz ifadeleri modellenmektedir. Yüz davranışlarının zaman-uzamsal gösterimi, kısa video parçalarından üç boyutlu artık ağ aracılığıyla elde edilmekte ve MIL torbalarındaki örnek olarak kullanılmaktadır. Hedefimiz, kısa sürelerde gösterilen ayırt edici yüz özelliklerini ortaya çıkararak anlaşmazlıkları tespit edebilmektir. Bu amaçla, olumlu deneyim ve anlaşmazlık oturumlarındaki örneklerin gösterimleri, derin metrik öğrenme ile daha ayrılabilir olacak şekilde eniyilenmiştir. Ayrıca, daha güvenilir bir ikili etkileşim analizi için etkileşimdeki bireylerin yüz ifadeleri birlikte incelenmiştir.

Deneylerimiz, yaklaşımımızın %71'lik bir başarıma ulaştığını göstermektedir. Birçok dayanak modelle karşılaştırmanın yanı sıra, aynı sınıflandırma problemi için altı katılımcıyla bir insan değerlendirme çalışması da gerçekleştirilmiştir. Yaklaşımımız, insanlardan %5 daha doğru tahminler sergilemekle birlikte tüm dayanak modellerden de iyi çalışmaktadır. Deneysel sonuçların önerdiği gibi, yüz davranışlarının güvenilir şekilde modellenmesi, ikili etkileşimlerin incelenmesine büyük bir katkıda bulunarak insanlardan daha iyi bir başarım sağlayabilmektedir.

Anahtar sözcükler: İkili Etkileşim, Davranış Analizi, Yüz İfadeleri, Çoklu Örnekle Öğrenme, Metrik Öğrenme, Derin Öğrenme.

Acknowledgement

First of all, I would like to express my sincerest gratitude to my advisor, Asst. Prof. Dr. Hamdi Dibeklioğlu. I felt his great support at every step in this long and tough journey, not only as an academical guide, but also as a friend. He knew how to motivate me at hard times, he was always ready to help and always stood by me. It has been a great pleasure for me to be his student.

I would like to thank Asst. Prof. Dr. Shervin Rahimzadeh Arashloo and Prof. Dr. Albert Ali Salah for being members of my thesis committee and their valuable feedbacks.

This master's thesis analyzes video recordings of face-to-face interactions collected as part of the project funded by The Scientific and Technological Research Council of Turkey through the 1001 Scientific and Technological Research Projects Funding Program (TÜBİTAK 118K162). In addition, I would like to thank TÜBİTAK for providing me financial support during my master's studies with 2210 program.

I wish to thank Burak Mandıra, Diyala Erekat, Oğuzhan Çalıkkasap and all other friends in Bilkent University for making my time here enjoyable, the sleepless nights more tolerable and for being great companions throughout our mutual thesis and education period.

I would like to express my special thanks to my dearest friends, Can Ufuk Ertenli and Berkay Karacaer, who were always there for me with their invaluable friendship and support. Many thanks to all others whom I cannot name here.

Last but not least, my deepest gratitude to my mother and father for all the things they have done for me. I dedicate this thesis to my dear sister, Defne, hoping the happiest life for her.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Communication is a complex phenomenon which is being researched for a long time in different -both social and biological- domains like psychology [1], psychiatry [2], medicine [3], physiology [4] and many more [5]. With its remarkable evolution, computer science became one of the most critical areas that scientists from various domains get help from. In the earlier stages, it was mostly in the form of signal processing, but with emerging deep learning techniques, now we observe audio or visual analysis of verbal, nonverbal communications in dyadic or crowd environments. Dyadic communications emerge in various cases, too. Doctor-patient, infant-mother, teacher-student, people who just met, or people in a romantic relationship; different kinds of communications have attracted the attention of researchers. By investigating those communications, we also aim to obtain various outcomes that we can benefit from. Understanding how we communicate and how we can improve our general approach to other people, what we look first at the other side of the communication or what kinds of reactions we give when feeling a particular emotion would open paths to brand new questions in the communication domain. Each step forward in this research domain and its applications in the real world would lead to a more effective civilization with better information flows, more empathetic people and hence, a happier environment, thanks to the well constructed communicative channels.

In order to contribute to the enhancements in this area, with this thesis, we consider dyadic interactions happening between romantic couples in a controlled environment. People would probably behave differently while they are communicating with their romantic partner, compared to the situations they are interacting with acquaintances, friends or family. Although we infer some general outcomes about dyadic communications from this thesis, as a historically interesting topic, we believe that obtaining clues about interactions in a romantic relationship would motivate other researchers that might be interested in these kinds of domains to further develop models like ours and increase the interest on the intersection of computer science and romantic relationships.

In the technical side, this is a video processing task with a major concentration on computer vision applied on human faces, utilizing several deep learning methods to predict the type of communication topic between couples as being positive or conflict, as they are told to speak on by the researchers from the Bilkent University Psychology Department where a unique dataset is collected. Since this dataset is used in a computational setting for the first time, there are no previous results to compare with. However, we have collected predictions from a group of human evaluators which might provide valuable information for comparison, along with the baseline results we report.

1.1 Research Question and the Challenge

In machine learning, there is a huge variety of data; text in natural language processing, voice, image, video, time-series like financial data or telecommunication signals, different forms of biomedical data are some of the popular ones. Likewise, labeling is not done only in one direction. Supervised, unsupervised, semi-supervised, one-shot, zero-shot, classification, regression... there are many different challenges to solve in the wild, and hence, many different appropriate ways to label the corresponding data. There are also many different sources of labels to be used for the dataset we use such as self report scores for different

sessions, expert and non-expert annotations, session labels as positive or conflict etc., which will be explained in Section 3.1 in detail. However, when we introduce subjectivity into the labeling, sometimes we can experience confusions. Let's take the trivial example of cat-dog classification into consideration. For an annotator, it is certain that whether the subject is a cat or dog. Or in a financial exchange data, it is for sure that the price was that specific value at that moment. This is generally not the case in most of the psychological data. 10 professional independent-coders working on the annotation of the dataset sometimes had such different opinions on the participants so that in one specific case, we have one annotator rating a person 7/7/6 (out of 7) for understanding, appreciation and care aspects, respectively, watching the participant's video, where one another rated the exact same person as 1/1/1 watching the exact same footage. They thought nearly the perfect opposites of each other, and this is only one example, there are many others like this one. So, in general, we can say that there is a considerable amount of variation between the annotations.

One other labeling that we have in our hands is self-reported values for the same questions (details in Section 3.1). This part could be even more problematic for some applications, since the participants do not follow a specific plan while giving the scores. At least we can assume that the professional annotators have a consistency within themselves personally, if not as a group. Some participants might give much lower scores than the others, for arguably much more positive sessions. Hence, we might think of these labels as supportive labels, rather than absolute truths. When we continue with these labels, we will introduce an unavoidable uncertainty coming from subjectivity. Our models will learn assuming these are absolute truths, and they will try to optimize itself themselves according to the possibly fluctuating labels, which might decrease the performance of the models.

We decide to define our task and labels as independent and objective as possible, in order to obtain results while avoiding the confusions stemming from subjectivity of the participants and the expert annotation. It is not that these labels are useless, but we want to keep the purity and simplicity at the maximum level. To this end, we mainly go with the binary classification of positive and

conflict sessions. Even if the participants may act more positive in their conflict sessions than some other couple's positive sessions; this way, we would guarantee that we have the *exact truth* of what they are told to be talking about, decreasing the effect of possible mislabeling.

Therefore, we can set our primary research question: *"Can we differentiate a dyadic communication which belongs to a positive interaction from another one that is based on a conflict happened between the pair, just by investigating their facial expressions?"*. In real life, we believe that there may be many use cases related to this question:

- This might be used in the security area in crowded places. In a setup close to this (two or more people are talking to each other face to face, not necessarily sitting at a table), with necessary improvements, a real-time argument detector can be made using the proposed classification technique. With "conflict" prediction, some departments might be alarmed if they would like to. In this setup, voice features might not be available in loud places like malls, casinos, concerts etc. and all we can reach may be the visual features.

- Other than already-romantic couples, this model can be useful when determining whether a date is going successfully or not. The trained models may still be a bit successful, but in these kinds of more-than-subtle changes in the topic of the dataset, the model would benefit from some more examples from the specific topic and a finetuning for those examples.

- Sometimes we might want to learn whether a meeting has been successful or not. If it is necessary, changing the labels to "success - not success", we can use the same approach and get the results. The possible use cases can include job interviews or business negotiations.

- People that do not speak the same language would be communicating at some places, with nonverbal interactions probably. In a scenario where we would like gain an insight about their communication, facial cues would play a great role.

- Parents might want to see whether the babysitter and their children get along well. Using an in-house camera system like modern babycams, the model can provide the information of the environment warmth. A multi-class (boredom, fun, anger, sadness etc.) classification with likelihood outputs can be produced as further improvements. The use cases can be further expanded where a nonverbal social interaction is present.

Even if we set our labels to be "positive" and "conflict" in a binary classification setup, there are still some challenges with the current labels. In some cases, participants have many much joyful moments together, laughing and joking, even during their conflict session; whereas some participants tend to be more calm and unexcited during their positive sessions, which makes them look like they are conflicting on a topic. We do not concentrate on comparing the two sessions and decide which one is positive and which one is conflict; rather, we consider each session (as pairs or as individuals) to be stand-alone data instances and try to decide whether this specific instance is coming from a conflict session or positive session. This way, predicting both sessions as conflict or both as positive is theoretically possible, rather than one positive and one conflict predictions. This increases the difficulty level of our task, since the question "which session is *more* positive?" is easier to answer, because it includes a comparison. In that case, even if the couple is more positive or negative than average, the model would now know exactly one of the footages would be negative, and evaluate the videos in this direction.

1.2 Related Work

In this section, we overview past studies related to dyadic interactions and the techniques used to investigate them. However, to the best of our knowledge, we are the first ones to work on automated facial analysis of dyadic interactions

between couples in romantic relationships. For that reason, we also provide information from studies, which concentrates on slightly different types of communications. Furthermore, our designed approach utilizes deep multiple instances learning, using frames from video segments. We included techniques similar to ours, along with an overview of the literature in the intersection of communications and interactions domains with computer science, including signal processing solutions, head and body postures, motion analyses and many more. We both focus on the psychological aspects and the technical solutions of several problems from this area.

Multiple instance learning, which is the main ingredient of our proposed model, is widely used in the computer vision domain. Applications on images include the study from Wu *et al.* [6], in which they use Joint Deep MIL to classify images, where the instances are both selected from the keywords for the images to be fed to a deep neural networks and the object proposals which are essentially the building blocks of the object of interest. Cinbis *et al.* [7] use multi-fold MIL where they iteratively train the object detector to infer the object locations.

In the video analysis domain, [8] classifies videos as anomalous or normal, by dividing the videos into fixed number of segments. 3D convolution features extracted by each of these segments would be the instances obtained in the bag formations in the beginning. Afterwards they use metric learning to separate the positive and negative instances. They reach 75.41% AUC in binary classification results. Our study is based on their research, with additional components like the use of two different views, employing a different representation, and adaptation to dyadic interaction. In another study, Phan *et al.* [9], divide the video into segments which becomes become the instances of MIL and look for specific actions in each segment. They quantize the instance-event similarity in a multi-class classification task and using this method they outperform 4 other baselines on a dataset with multiple complex event videos. Medical image and video analysis (MIVA) tasks represent another domain in which multiple instance learning is used. As explained in the survey from Quellec *et al.* [10], in this domain, researchers set their bags as a video or image. If videos are used, they extract features from a segment and using the globally assigned diagnosis labels, many

of them conduct classifications. As we can see, multiple instance learning can be widely used where we have a label for a whole (e.g., video), but the clues are hidden in its sub-parts (e.g., video segments).

We use multiple instance learning to extract information from videos of dyadic interactions. These interactions are widely studied by the researchers both from the computer science and psychology domains. We can infer a lot about the relations of interactants from the intersection of these domains.

There are important concepts worth mentioning when investigating the literature on dyadic interaction. The action of imitating someone, or mimicry, is one of them. Wu *et al.* [11] find connections between the nonverbal mimicry and person's communication skills, where they automatically detect mimicry between medical students and volunteers who act as their patients. Cheung *et al.* [12] present support for facial mimicry serving to promote interpersonal rapport, which can be defined as understanding the others' feelings or ideas of other parties of a communication. [13] concentrates on the relationship of mimicry and cognitive and emotional empathy, where they exhibit findings showing that emotional and cognitive empathy differences among different individuals are associated with the level of facial mimicry. Emotional contagion (primitive) is another important concept, which is defined as the tendency of people to mimic and synchronize their multimodal behavior during interactions and, consequently, to emotionally converge to each other. Varni *et al.* [14] investigate both facial expressions and conduct sentimental analysis to capture emotional contagion in unimodal, multimodal and cross-modal levels at the same time.

There is a well-known concept called *chameleon effect* in psychology [15], which refers to unconscious mirroring of expressions, postures, and other physical and verbal behaviors of participants that get along well with each other in a dyadic conversation. The importance of chameleon effect in a dyadic relation is that it facilitates creating rapport with the one who is being mirrored, as it is found that people feel more connected with others that behave alike [16, 17, 18]. For instance, several studies show that synchronous nonverbal behaviors between teacher/student dyads not only create rapport between the interactants, but also

boost the learning performances of students [19, 20, 21].

The verbal or nonverbal interactions of 2 or more people is being researched in the computer science area as well as their implications on the psychology domain. Psychological researchers consult computer vision techniques in many cases while studying dyadic interactions. In general, according to authors' findings in [22], while making judgments, the nonverbal cues can provide rich information about the person of interest. This factor led many researchers to obtain information from nonverbal communicative channels and to use them in their studies to reach various goals.

The important concept of social signal is explicitly defined as any action or overt behavior, regardless of its form, intent, or the performer's awareness, that is carried out in the presence of another person [23, 24]. Social signals extracted from different modalities, such as mutual gaze, body posture, interpersonal distance, vocal behavior, and hand gestures, are employed in the analysis of nonverbal interactions. Social signal processing, i.e., the computational analysis of nonverbal behavior aiming at bringing social intelligence in computers [25, 26], increasingly attracts attention of interdisciplinary researchers. Vinciarelli *et al.* [27] argue that machines should be more aware of these signals and obtain social intelligence to be more efficient in socially aware computing domain.

Dyadic or group interactions and their computational analyses can be seen in a wide variety of applications [28]. In [29], authors investigate the interactions in collaborative learning environments, integrating techniques from computational psychometrics and deep models like CNNs. By investigating human behavior in such environments, they identify evidence about teamwork skills. In another study, authors collect a dataset, where they investigate the group interactions in a board game playing scenario [30]. Patient-doctor interactions is another application area. In [31], the authors show that the patient's affect can be estimated by taking the doctor's affective cues into account during their interaction. Interestingly, their linguistic results outperformed facial analysis for most of the affects.

There are many different approaches investigating dyadic interactions, including the use of wearable sensors [32, 33], ECG [34] and EEG signals [4]. Pose related visual features like head movement [35], head posture (e.g., facing straight, facing down), arm posture (e.g., far from the body, touching the head or face) [36], and full body pose [37] are also employed in the analysis. In addition, verbal interactions are examined using language models [38].

To find the effects of synchrony in dyadic interactions like psychotherapies, and its influence on perceived empathy and rapport, authors utilize Motion Energy Images (MEI) analyzing the motion in the regions of interest (ROI) [39]. They choose ROIs as the interacting people and create two continuous time series for further processing. In a follow-up study [40], Tschacher *et al.* find strong effect sizes for synchrony to occur, applying automated motion energy analysis.

Cerekovic *et al.* [41] analyze human-agent interactions looking at the personality of the subject as well as the nonverbal behaviors observed during the interaction with the agent. They concentrate on body and head pose, hand activity, and visual appearance of the subject, as well as modeling vocal cues. In [42], authors concentrate on head movement patterns like angular displacement and angular velocity between the infants and their mothers. They find out that investigating head movement can give important clues on understanding the apparent emotion and interpersonal coordination, as the head movement is strongly correlated with the interactive context between the infant and the mother according to the results.

Hagad *et al.* [36] design a model to predict rapport of dyadic interactions automatically, using the head and arm related postures and congruence. In their model, Histogram of Oriented Gradient (HOG) features are employed to classify postures using Support Vector Machines (SVM). To predict if there is a rapport, posture congruence is determined depending on the similarity of the participants' postures, i.e., postures are said to be congruent when participants hold their bodies in the same position as each other. Terven *et al.* [33] use smart glasses that contain high-definition cameras in the bridge connecting the two lenses to detect nodding and therefore mirroring noddings between the dyads. They extract

facial features using Active Appearance Models [43] and apply stabilization on the extracted features which are later used for head gesture modeling via Hidden Markov Models (HMMs). In another study concentrating on facial expressions [44], authors present a database where a group of people are interacting with each other. They use deep models processing facial action units while providing baseline results for their database. These studies suggest that concentrating on head and face features in particular, is found to be beneficial while investigating different aspects of social interactions. In [45], authors investigate the relation between facial expressions and rapport establishment using an automatic facial expression coding tool called CERT. They employ facial action units analyzing dyadic interactions in different setups.

Facial expressions in dyadic interactions have also been a topic in the research on Generative Adversarial Networks (GANs). In [46], Huang and Khan propose DyadGAN, which models the effect of one party of a dyad on the other, concentrating on facial expressions. They use Emotient's Facet SDK [47], which extracts 8D facial expression descriptor vectors, representing the likelihoods of emotions such as joy, anger, surprise, fear, contempt, disgust, sadness and neutrality. [48] proposes another GAN model that would learn semantically meaningful facial expressions employing Conditional LSTMs.

Two class classification is a common task when investigating dyadic or multiple participant interactions. The classes are generally selected as positive and negative, which we can map to the positive experience and conflict cases in this thesis, respectively. In [32], by using the individual behavior features such as movement, location, audio and being face-to-face (using infrared sensors) obtained from wearable social sensors as well as examining the social interactions in a team with questionnaires, authors aim to classify the participants' affect states and group cohesiveness as positive and negative. In [35], authors try to classify behavior codes such as Acceptance, Blame, Positive and Negative with classes "high" and "low", indicating the codes' presence levels by utilizing head motion patterns. [38] is another study that classifies the behavior codes (e.g., anger) as "high" and "low" using language models. The dataset they use resembles the employed dataset in this thesis, as there are 134 real life couples attending

marital therapy, who are talking over 10 minutes in different sessions. Authors use the top 20% positive and the top 20% negative instances for modeling and classification, focusing only on the extreme cases.

Most of the video datasets, especially the ones that are collected in a controlled setup, do not contain large amount of data. For instance, both of the Group Formation Task dataset [44] and the personality analysis dataset collected in a recent study of ours [49], include only 60 participants. Datasets like Kinetics-700 [50] or First Impressions [51] include video clips that are obtained from different YouTube videos and this makes it easier to create a large-scale dataset. In this study, we use a recently collected romantic relationship dataset [52] (will be referred to as Romantic Relationship Dataset in the remainder of this thesis), which includes 167 couples (334 subjects).

Due to the limited number of data samples in the dataset, we use cross validation for optimization in order to benefit from the whole dataset. Cross validation can be defined as subject independent, or subject dependent. It can be observed that the utilization of subject independent cross validation might reduce the success rates compared to subject dependent case, since in the subject dependent cross validation case, models explicitly learn the behavior of a test subject. Many studies including [4], [53] and [54], report lower accuracy in the case of using subject independent cross validation (compared to the subject dependent one). Hence, we can conclude that subject independent cross validation is a harder and more realistic approach in general.

Chapter 2

Classification of Dyadic Interactions

In this chapter, we describe the proposed architectures and models and the general workflow of the classification of dyadic interactions between romantic couples. The dataset will be explained in detail in Section 3.1, but it is worth mentioning that we have 3 cameras (2 frontal and 1 side views) recording 2 different 10 minute sessions, namely the *positive* and *conflict* sessions which we conduct binary classification on. In conflict session, couples are discussing on a topic they have conflicts on, whereas in positive session, they talk about positive past experiences they share. While studying the visual domain which we try to extract information from, first, we process frontal view video frames to extract some visual features like facial landmarks and action units. Using those 2D landmarks, we normalize and warp the faces such that we obtain normalized frames which includes only the faces of the participants standing as straight as possible. Then, in order to strengthen the connections between the frames and not being lost in a huge video, we divide each video into shorter periods. No matter what type of model we use, we have to get the best out of an approximately 10-minute-long video. This led us to use videos in segments in many different experiments. We propose a solution adapted by the anomaly study of Sultani et al. [8] to our dyadic interaction case, where we feed these shorter videos to a network which combines

Multiple Instance Learning and Metric Learning. The details of the techniques used and the process are explained in the following sections of this chapter.

2.1 Face Alignment

Using normalization is a crucial step when dealing with faces, as in nearly all other types of data in machine learning. The trivial approach would be detecting the face and cropping a sub-image around it for further usage. However, with this basic technique, the faces would not be aligned due to head shape and pose changes and the models might have trouble detecting subtle changes in mimics. In order to eliminate this problem, we warp the faces using some critical points in the faces, namely the 2D landmarks, in order to normalize each frame in terms of translation, rotation and scale. Figure 2.1 showing the OpenFace features including those 2D landmarks is directly taken from OpenFace 2.0 study [55], since we cannot disclose the participants from our database with features applied on their faces.

We obtain the frames from the videos using the ffmpeg package. The frequency of frame extraction is set to 13.6 after investigating the average fps values of the videos, which are varying (explained in more details in Section 3.1). We select the frontal views to process, since we have a much more clear vision of the faces and in a physically less active setup like this, we might extract most of the information from the participants' faces and mimics. If we were to concentrate on the body pose and movements, side view might have been a better option. Feeding those frames to OpenFace models, we extract many facial features including the action units (we use 17 dimensional action unit intensity model only) and 2D landmarks. After removing the global rigid transformations such as translation, rotation and scale from the 2D landmarks, we pass to the face warping stage. In order to do that, we use piecewise linear warping in which we transfer the landmark coordinates onto their original normalized locations. The output is scaled and cropped such that only the face with a black background is left, where landmarks of all faces among the database are residing in the exact same location.

The resulting images might seem less informing to human eye, but since they are pixel-to-pixel comparable across the database it would be much more easier for our models to detect even the most subtle clues happening in our faces. The resulting normalized faces has a resolution of $512 \times 512 \times 3$.

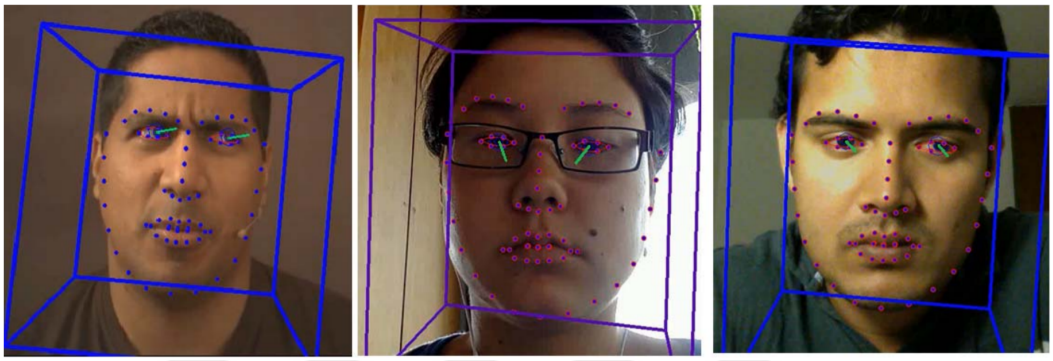


Figure 2.1: Some features obtained from OpenFace such as head pose, gaze, eye landmarks, 2D face landmarks, 3D bounding box

If the normalization step is somehow problematic, since it is generally the very first step of all the calculations, this error might affect the whole process exponentially. After all, not having clean data can be a real problem in today's deep learning challenges. Although this step is straightforward for many tasks, in a setup like this, there happens to be many difficulties with the subject. Since the participants are not totally concentrated on being recorded at the moment of their speech, they do not look directly in the camera or stay still with their heads and bodies. Due to the camera angles and the participants' head directions, sometimes we do not have the access to the whole face. OpenFace models also struggle with these kinds of frames and 2D facial landmarks may not be perfectly accurate at times, which affects the face warping procedure. In addition to head pose and angle imperfections, we observe occlusions like eyewear, beard, accessories or participant's hands or arms sometimes, which causes distortions in the landmark detection. There are some different cases when participants' faces are completely lost on the frame and we are left with -expectedly- random normalized face images which does not include any face at all. In order to capture the full motions, bodies and the faces, the cameras are not placed perpendicular

to the participants, but rather they are placed at a (small) angle. So, most of the times, the distortions in the normalized faces are stemming from unseen right part of the faces, especially if the participants tend to tilt their heads that way. Consequently, although the OpenFace model gives nearly perfect landmark locations and the warping algorithm works completely fine, we are obliged to face a non-perfect normalization in data, which makes the task harder to solve.

2.2 Spatio-temporal Representation

In order to extract spatio-temporal representations as feature maps from the video sequences, we use the 18 layer ResNet-3D architecture in [56]. Although the authors in [56] proposed a model with slight improvements in accuracies, ResNet(2+1)D, which divides 3D convolutions into two separate and successive operations (a 2D spatial convolution and a 1D temporal convolution); we select ResNet-3D architecture to benefit from its memory advantages. Actually, these models are created and tested for action recognition tasks on datasets like Kinetics-400 [57] and Sports-1M [58], whereas our task is not directly an action recognition task. Our dyadic interaction and action recognition tasks resemble each other in many ways, as there is an effective motion in the video that needs to be analyzed in a spatio-temporal way; but in our case we do not want to understand *what* is the action -since we know that it is someone speaking-, we actually want to understand what is the *context* of that repetitive action throughout the whole dataset. So, we do not need to go into details of action recognition by using the much more complex models and their last layers which are specified to find the action classes. We rather need to grasp the idea of *how* to investigate the action, which is well provided with the low/middle levels of the architecture. Hence, ResNet-3D model with the original 3D convolutions might even be more suitable for our case. This way, we would eliminate the overspecification problem and get more generalizable, well representative feature vectors as well as coping with the memory restrictions.

In Resnet3D-18, only vanilla residual blocks are used without the bottlenecks. The video clip of any size is passed through several 3D-convolutional layers. In the layer represented with green in Figure 2.2, there is a Conv3D with kernel size of (3, 7, 7), followed by a 3D batch normalization layer and a ReLU activation. The red layers are consisting of the same schema, but they have a kernel size of (3, 3, 3) this time. The purple layer which represents the downsampling, is another Conv3D with (1, 1, 1) shaped kernels. ReLU activation does not exist after downsampling layers. After all Conv3D layers, there exists an adaptive average pooling layer, which practically compresses the 3D video representation into a 1D vector by taking the average over the entire spatio-temporal volume. In the original model, there exists a fully connected layer with 400 output dimensions, representing the number of classes in Kinetics-400 dataset.

We use transfer learning in the feature extraction process. The ResNet-3D model we use is the one pretrained on Kinetics-400 [57] dataset. We take that pretrained model, freeze the weights, and extract the 512D feature vector which is the penultimate layer. Instead of extracting frozen features from it, we could have finetuned ResNet-3D as an end-to-end model in the architectures in which we use the features. However, even the feature extraction process is extremely heavy in terms of time and trying to finetune the networks to extract features more specified for our dyadic interaction classification task would be a real burden for our network. Considering that we have so many folds and hyperparameter sets which we talk in detail in Chapter 3, and the fact that these are all independent variables and increase the experiment count exponentially, we keep the ResNet-3D weights frozen and separated the feature extraction and training stages. In the end, we have multiple 512D feature vectors in our hands for a single video clip, with the number of feature vectors changing according to the technique we use (fixed period or fixed partitions, which will be explained in the next paragraphs).

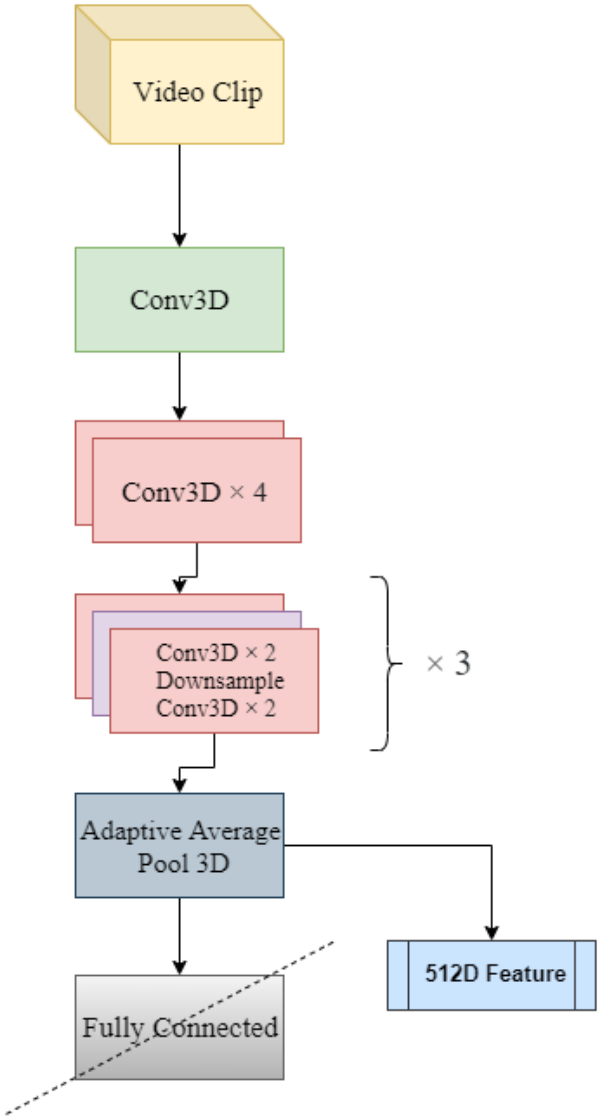


Figure 2.2: ResNet-3D-18 architecture

With an approximate length of 10 minutes, the average video is consisting of 9000 frames. Each frame has 1080×1920 pixels for RGB channels. In some setups, we consider multiple videos as the input, for one pass in our network. This makes our input significantly large in size, so we need to find a way to efficiently use the video clips.

To be used in many of our experiments, we extract 1D vectorial features from

the videos as their smaller-in-size representations. Only one vector, even possibly with a huge size, might not be very informative of the whole video, just because of the reason we have stated in the previous paragraph - videos contain too much information. Furthermore, in a wild-case scenario there might be an instance of the same kind of video for only 15 seconds, whereas we might want to investigate a 1 hour version of the same video, too. The same-length vector might not be the best way to represent these two instances at the same time, because of the varying information depth. In order to handle this issue, while extracting the feature maps, we divide videos in particular ways and consider those much smaller video portions as individual clips to be fed to the networks.

There are mainly two techniques that we use: fixed partitions and fixed periods. In the fixed partitions method, we select the number of partitions we want, divide the total frame count with that number and use that many frames for one partition. In the end, we are left with the same amount of sub-video representations for all of the videos. The other method, fixed period, is fixing the period as its name suggests. There are two parameters for this feature extraction method: window size and step size. Window size determines how many frames (hence, seconds) we are investigating and the step size tunes the overlap between the windows. For example, if we take 100 as our window size and 40 as the step size, the first feature vector will be extracted from the frames [0, 100] and the next vector will be from [40, 140], resulting in 60% overlap between each partition. One benefit of using overlapping windows can be saving ourselves from leaving an important part of the conversation in two different partitions. However, using overlaps would significantly increase the need for resources like training time and memory. In [53], authors conclude that although using overlapping-sliding-window technique is beneficial for subject dependent cross validation, when subject independent cross validation is used, this technique does not improve the results. It only introduces the resource requirement increase.

We explain forming train-validation-test splits and folds further in Chapter 3, but we continue with subject independent cross validation, hence the data-subsets and folds are created such that one data sample (a participant dual with both conflict and positive videos) is only appearing in the test set once, without any

overlaps. So, the model trained for that specific fold does not see any data from that participant dual in either train or validation sets. In this way we eliminate the chance to overfit at some individuals that are seen in the train set. Since we have only 167 participants but in total 60 minute of videos the process for each one, the chances are arguably high that the model would perfectly predict the ones that it has seen in the train set. As we choose to continue with subject independent cross validation, not using overlapping sliding windows would be our first choice.

As mentioned earlier in this section, there are two main video dividing techniques. Since we decide not to use overlapping sliding windows as our main approach, we are left with the non-overlapping fixed partitions and fixed periods techniques. If the video lengths are greatly different than each other, using fixed partitions method would result in feature vectors, coming from a high variety length of videos. This would not be that desirable, since we would like to have similar representations for each video part. If the videos are different in lengths, one feature can represent, let's say, 1 second, and another one can represent 1 minute, in which our pretrained model might concentrate on different aspects. We would need signs from a video partition to tell us that it is a video from a conflict or positive session, where the duration of the action of interest would probably matters. Hence, it would be better to extract features from similar length videos, if not equal. For our case, since all of the videos are captured in a supervised environment and the experiment duration is fixed, all videos have similar numbers of frames. So, fixed period and fixed partitions would not change things a lot. When we use fixed period approach, since the videos are not exactly equal in duration to each other, there may be 2 to 3 more features extract for some video clips. This makes the amount of conflict features and the positive features unequal and this is the downside of using fixed period technique. The disadvantage of the other technique, fixed partitions, is that the features are coming from different length videos, but the length differences are less than 1 second. In comparison, having 50% conflict and 50% positive features seems to be a slightly better advantage, since the duration difference in fixed partitions method is more negligible.

2.3 Multiple Instance Learning

In most of the classification tasks, the instances are directly correlated with its assigned label. For example, in a classical cat-dog classifier, if the image is a "dog", it means that the contribution from every pixel makes that image a dog. However, in some cases like ours, the data instances might be comprised of several pieces that might not be assigned to the same label. If we restrict the case into the video domain, this generally happens when the video is weakly labeled, that is, there are video-level labels which are annotated due to some smaller part or parts in the video. This is also the case in [8], where they are trying to detect anomalies in surveillance videos. In a surveillance video, the anomalies do not always happen, but if it exists even for a short period, the whole video would be labeled as an anomalous video. Since they do not have access to the exact period of the anomalous event due to the weakly labeled videos, they apply Multiple Instance Learning (MIL) rather than obtaining accurate segment-level labels. Our case is not exactly the same as theirs, but both tasks have a lot in common. The main similarity is probably the fact that a human annotator would consider a short instance that would convince him or her that if there exists a anomaly or a sign of conflict, while annotating the whole video. Not all of the frames would have equal contribution during the annotation, the annotator would look for a "moment of interest". So, the main similarity would be that we both are trying to catch an interesting part of the input in order to predict that it is conflict or anomalous. However, one strong difference would be that the participants are expected to be talking on a conflicting case throughout the session, while the anomalous event would not exist at all for most of the times. If we were to distinguish a severe fight between the couples and a totally positive session, our case might not have needed this kind of setup. By utilizing MIL with dividing videos in smaller parts, our main assumptions can be summarized as follows:

- In positive videos, whole session would probably be in a good mood, without drastic changes in the actions of the participants.
- In conflict videos, we observe that many couples do not fight on a topic all

the time. They rather laugh, smile and seem happy at times. This might be due to the fact that they are told to talk on a topic they have conflicts on, but the disagreement might not be so strong or recent that it would not be effective to change their mood.

- Therefore, the arguments might happen occasionally, which we can correlate to the "actions of interest".
- The short sequences of videos from both positive and conflict sessions, might be more positive or more conflicting individually. That is, a subsequence of a conflict video might be more positive than the actual positive video, since the participants change attitude during the videos. The best we can do is to assume that most of the conflict subsequences are conflicting videos, and the others are positive videos, since we do not have frame or sequence level annotation.
- If we detect a sequence of strong conflict anywhere in the video, we might succeed by predicting that this is a conflict session.

This is the point where we benefit from MIL. We represent the positive and conflict videos as "bags", and the instances in those bags would be the video segments that we form. In the positive bag, we would assume that the *most* conflicting video segment is not enough for us to call that video a conflict case, whereas there will be at least one instance in the conflict bag, that would indicate the whole bag belongs to the conflict class. In simpler words, when comparing two bags, if the maximum value in a bag, is larger than another bag's maximum value, then we will call the bag with larger value a *more* conflict bag, i.e. video. Hinge loss in the form of Eqn 2.1 can be used to optimize the MIL objective function, where BI represents bag instances.

$$\max(0, 1 - \max(BI)) \tag{2.1}$$

While using this formula, we want the instances of positive and conflict classes to be as far as possible, and we use MIL Ranking model with metric learning,

adapted from [8], in order to do that. Metric learning in general, tries to discriminate the inputs by their similarities according to a distance metric. Deep metric learning, which uses neural networks to learn how to discriminate the input features according to its distance metric, is used in various computer vision tasks such as face recognition, face verification, anomaly detection and 3D modelling. For example in face verification task, a "correct" person might have different poses, different facial expressions, occlusions, accessories etc. but the model needs to verify that he or she is the correct person most of the times, if not always. Here, the metric learner model would aim to decrease the distance between the images of the same person, while increasing the distance with the other people. As this example shows, this technique is well applicable in areas where we want to discriminate any type of input, that would probably have different variations of its own, but still belongs to the same class.

In our case, we have 512D features extracted from the subsections of a video, which would be the instances in the bags. If we can discriminate the positive and conflict bag instances by increasing the distance between their representations, we can classify the videos by looking at those sub-video separations. In our approach, where we combine multiple instance learning with metric learning, we are trying to discriminate the instances in the bag that is formed in the MIL part. We use the modified version of Eqn 2.1 in order to complete the positive-conflict discrimination as follows:

$$L_D = \max(0, 1 - \max(BI_{conf}) + \max(BI_{pos})) \tag{2.2}$$

In this equation, we can clearly see that as the maximum valued conflict instance gets larger and the positive get smaller, the loss decreases. Then, as they also do in [8], adding a temporal smoothness factor to this equation would be meaningful, in order to prevent extreme changes in scores between adjacent video segments. The conflict behavior would not change instantaneously, but it would rather show itself and disappear in a period. There is also a sparsity constraint as a regulator to the scores. In the end, the total equation would be like Eqn 2.3, where t represents time as in the video segments:

$$L = L_D + \lambda_1 \sum_t ((BI^t_{conf} - BI^{t+1}_{conf})^2) + \lambda_2 \sum_t (BI^t_{conf}) \tag{2.3}$$

2.4 Loss

Other than the loss that we define for the MIL case, we try several other loss functions in different experiments. Our main challenge is a binary classification task and there are multiple options to use, when it comes that. Some binary classification tasks are discrete, that is, the classes have no *amount* of belonging to that class. They are either in or out of the class. Although what the participants are told is as discrete as it gets (positive-conflict), the positivity of the discussion might fit into the continuous domain better. Although the most common evaluation metric is accuracy as in any classification task, they are all trained using some kind of continuous function as the loss metric.

The binary cross entropy loss (BCE) or its combination with sigmoid layer, called "BCE with Logits Loss", are widely used in binary classification tasks. Hinge loss is another option in this task and it is the piecewise linear version of BCE in terms of graph similarity. We train some of our models using both of these losses while experimenting.

Margin Ranking Loss, as shown in Eqn. 2.4, is one other suitable loss function for this task. Originally, the objective of ranking losses is to predict relative distances between inputs, just as in metric learning.

$$L_{MR}(y_{pred}, y_{true}, \alpha) = max(0, \|\alpha - y_{pred}\|^2 - \|\alpha - y_{true}\|^2 + margin) \tag{2.4}$$

Although it is essentially a ranking loss with multiple classes, it is perfectly fit for our case with little alterations. Here, the variable α represents the *anchor* value, which we will take 0, since our labels are selected in the set {-1, 1} for conflict and positive cases, respectively, and zero centered. We would also consider

the L1 distance between the label and the prediction. The altered form can be found in Eqn. 2.5.

$$L_{MR}(y_{pred}, y_{true}) = max(0, \|y_{pred} - y_{true}\| - margin) \tag{2.5}$$

Here in Eqn. 2.5, we are looking for the distance between the prediction and the true label. If the distance is less than the margin that we have defined earlier, the loss would be 0. After the distance surpasses the margin value, the loss starts the increase linearly. Here, the selection of the margin value can be critical. For the label set {-1, 1}, let us assume that we use 1 as the margin. If the model predicts 0 for all of the instances in the dataset, it would always get 0 as the loss, without learning anything at all. The same applies for the margin values even larger than 1. In the range (0.5, 1), if the model fails to learn again, it would be much more logical for it to predict 0, since the punishment from the loss function, would not be that harsh in comparison with the case it randomly predicts a number in the range [-1, 1] and fails with a probability of 50% (fail here means being closer to the wrong label). Hence, we select our margin value to be 0.5 as a reasonable value for our label selection. Here, our main aim is to push our model to be as confident as possible, by surpassing the margin if it can. Of course, being "more" sure and predicting incorrectly will be punished more.

While evaluating the models optimized using margin ranking loss, we only look at the sign of the output. If the output is positive, we predict positive; if negative, we predict conflict. The same loss cannot be used while evaluating, because even if the model is unsure (for example predicting 0.01), it selects a class anyway and it is either correct or not. So 0-1 Loss is used for the evaluation, when margin ranking loss is applied during training.

Chapter 3

Experiments and Results

In this section, we share the experimental setup, the results from our experiments and discuss possible outcomes and reasons that we can infer from those. We also present a human evaluation study and compare the baseline results with the proposed method.

3.1 Romantic Relationship Dataset

The main dataset that we conduct our research on is collected by the researchers in Bilkent University Psychology department, under supervision of Dr. Gul Gunaydin. This dataset, which we will call *Romantic Relationship Dataset* [52], is consisting of 167 dyads (couples who already know each other and in a heterosexual romantic relationship). The experiment is consisting of two different stages, namely the "positive" and "conflict" sessions. As the names might suggest, in the positive session, the participants are requested to decide and talk about a good memory they share, an event that made them happy or anything "positive" that they would like to talk about them and enjoy. Contrarily, in the conflict session they are asked to decide and talk on a topic that they have argued about,

possibly an unsolved argument, or in general, a topic in which they have conflicting thoughts on. Before each session, the researcher who told what they expect from the participants explicitly indicated that they are expected to be talking naturally and relaxed as if they were talking in the outside world regularly, which is also recorded at the beginning of all the videos (the parts where the researcher speaks does not intersect the participants' talking). Here, the researchers from the psychology department aimed to record the enjoyable and conflicting discussions happening between romantic couples, noticing the differences between them, and also get the participants' ideas on these sessions along with the professional analysis from the psychologists' side.

In all experiments, participants started with a self report questionnaire about their partners, followed by the conflict session. They filled the questionnaire again, right after they finished with the conflict session. After that, they continued with the positive session and filled the same questionnaire, this time indicating their thoughts on the positive session instead. The results of the questionnaires contains perceived partner responsiveness (PPR) scores for each participant. PPR refers to the extent to which the participant feels their interaction partner cares for, understands, and appreciates them. So for each dyad, both the perceiver's (the person providing the PPR rating) and the partner's (the person whose PPR is rated by the perceiver) behaviors might be related to the perceiver's PPR scores (although it is expected that partner's behaviors would be more relevant). To measure PPR, the researchers asked participants to indicate their agreement with each statement below on a 7-point scale, where 1 indicated "Strongly Disagree" and 7 indicated "Strongly Agree":

- During the interaction I felt that my partner understood me.
- During the interaction I felt that my partner appreciated me
- During the interaction I felt that my partner cared about me

A team of independent coders consisting of 10 professionals, later on, evaluated these videos according to the same measures: understanding, appreciation

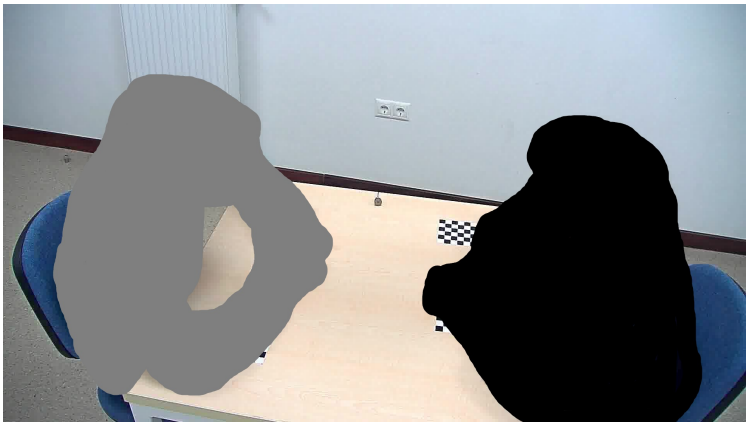
and care. For further utilization of this dataset, researchers will have individual scores on all 3 aspects in their hands, again on the same 7-point scale that the participants have filled beforehand.

There are 3 different cameras recording the sessions, all synchronized beforehand mechanically in order to be able to use the videos from different angles at the same time when needed, and be sure that the captures are from the exact same time point. As it can be seen in Figure 3.1, 2 cameras are facing the participants, and the third camera is located as it would record them from the side view, able to see both participants at the same time. For confidentiality reasons, we cannot share the exact images of participants in order not to disclose their identities. Here, the gray color represents the male, and black represents female. In order to be consistent along the dataset, women are always facing one camera, while the men are also doing the same. It would be beneficial to mention that the cameras facing the participants are not located in a 90 degrees direct position to the faces, but instead both nearly have 60 degrees angle with the faces, trying to prevent any occlusions stemming from one participant over the other. Although this angle makes one side of each participants' face a little bit unseen at some of the moments, this feature might have made the dataset much more alike the real life situations where we don't always have a direct look on the faces. However, the spot in which the participants sit, and the camera positions do not change along the dataset ever.

The videos are around 10 minutes for each session, with 3 cameras recording and 2 different sessions, it makes approximately 1 hour video to analyze per dyad, excluding the parts where the researcher tells the participants to start and finish, which lasts around 20 seconds in total per session. Each of these videos has 44100 samples per rate and 2 (stereo) channels as usual. Frame width and height are 1920 and 1080, respectively, resulting in a resolution of 1080p. However, fps (frame per second) values are not consistent among the whole dataset. There was a technical problem with the recording setup, while trying to record three 1080p synchronized videos, the fps could not be produced at the desired rates and it has dropped in between 13.5 - 14 for all videos. After sufficient investigation on several samples, we observed that the decrease in fps, is not due to some frame

(a)

(b)



(c)

Figure 3.1: Frontal views (a) and (b), showing female and male, respectively, and the side view (c)

drops at specific periods, but it is more like spread around the whole video. That is, if a video has 13.8 fps for example, it has 13.8 and alike values for each second in the video, with a low standard deviation (instead of 25 fps for most of the seconds and huge frame drops for some instances). While getting the frames out of the videos, we should be using a fixed fps value, since the same amount of frames should mean the exact same amount of seconds. We have fixed this fps value to 13.6 for all videos. There have been some missing frames for the videos with higher fps values, and some repetitive frames for the ones with lower fps values.

3.2 Experimental Setup

3.2.1 Model Training and Implementation Details

Before starting our experiments, we select which input feature technique we should use in our multiple instance learning model. Since we are using bags of video segments while implementing the proposed architecture, in order to have equal number of samples from each participant, we start off our experiments with the "fixed partitions" method here. The duration change for those partitions are not expected to have that much of an effect.

When we look at the number of video segments for the best hyperparamater combinations, we see that all 4 fixed partition counts occur as the best option in at least one fold. Although there is no number that we can directly select, we observe that the higher number of divisions are working slightly better, as they are more commonly found in the best performing model hyperparameter sets. That is, having 3-5 second long video clips are more likely to give better results than the clips that last for 7-8,5 seconds. Among those models which are fed with features that has more partitions, 120-partition option is very slightly better than the 210 one, but we cannot say that it is by far the best for all setups. Overall, we can conclude that approximately 5-second-long video clips seem to

be working fine for our task.

In Table 3.1, we show the considered intervals and best choices for the hyperparameters that we use during the MIL experiments. Here, warm-up epoch means that the amount of epochs need to pass until we start finetuning the ResNet-3D, if we apply finetuning. The initial epochs are for the hidden layer's "warm-up" and creating a better weight set than the initial weights. Skipped epoch is also used in some cases where the best model is chosen directly as the initial model. We prohibit early-stopping in the first N epochs in this setting, but the best models are the ones that do not utilize these functions anyway.

Table 3.1: List of Considered Hyperparameters of MLP for MIL

Hyperparameter	**Considered Interval**
Number of units in the 1st hidden layer	[64, 2048]
Number of units in the 2nd hidden layer	[16, 1024]
Number of units in the 3rd hidden layer	[4, 256]
Number of hidden layers	{1, 2, 3}
Initial Learning Rate	[1e-6, 1e-1]
Weight Decay	[1e-4, 1e-1]
Batch Size	[32, 64]
Dropout Rate	[0, 0.8]
Optimizer	{Adagrad, Adam}
Weight Initialization	Xavier Normal
Early Stop Epoch	{50, 100, 500}
Max Epoch	{5000, 30000}
Skipped Epochs	{0, 10, 35, 50, 100}
Warm-up Epochs*	[0,500]
Number of partitions	{75, 90, 120, 210}
Window size	{30, 40, 60, 100}
Step size	{20, 30, 40, 50}

*: Where ResNet-3D is finetuned.

3.2.2 Cross Validation

Since we do not have explicit train, validation and test sets, we use k-fold cross validation, in order to be able to evaluate every sample in the dataset as a test sample, while obtaining a good enough model with training samples as much as possible. To this end, we form 9 folds from the dataset with 167 instances with no subject overlaps, where instance means all the conflict and positive videos of a couple, resulting in 4 videos in total. We made sure that an instance is always found in only one fold and not any others. 5 of these folds have 19, whereas the other 4 have 18 instances. Due to the fact that 167 is a prime number, it was impossible to have folds that are consisting of the same number of data instances. The number of folds is chosen as 9 to provide much similar number of pairs in each fold. All the folds are fixed, and the same ones are used in the entire experiments. For all of the experiments, we have selected 7-1-1 train-validation-test split, as shown in Figure 3.2. Split-n has n^{th} fold as its validation set and $(n+1)^{\text{th}}$ as the test set. For example, for Split-8, if folds [1-7] are used for the train set, $\{8\}$ is the validation set and $\{0\}$ is the test set. The cross validation was for one level rather than two, which means we do not loop the validation set inside the train set. For one train-validation set combination, the hyperparameter set which gives the best validation scores is selected to be the optimal set for that specific split. Looping the validation set inside the train set would multiply every experiment we do by 8, which would be computationally very expensive. Different models generated from the same architecture for each 9 splits do have different optimal hyperparameter sets. With the model resulting in the best validation score, the test scores are obtained and reported in the further tables in this section. Furthermore, we will touch upon the indifference between test and validation results of individual folds coming from models trained with different splits while sharing Table 3.3 in the following pages.

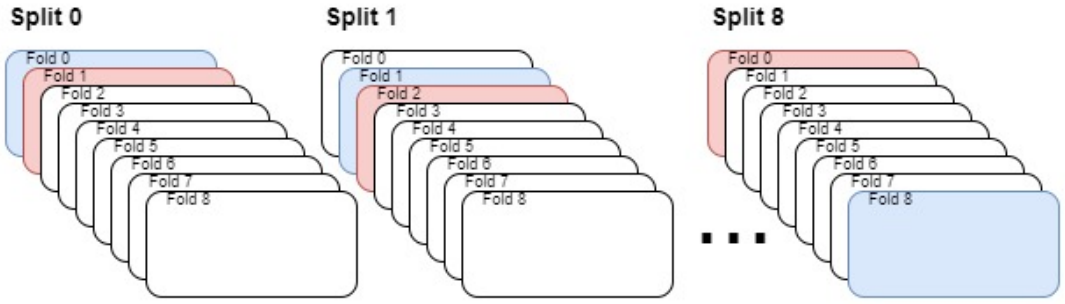


Figure 3.2: 9-fold cross validation and corresponding train, validation, test set formations. White, blue and red represents train, validation, test sets, respectively

3.2.3 Evaluation Metric

Using the proposed model, we obtain scores for each video segment that indicates its conflict level, and the maximum scored segment represents the video-level prediction. Different from usual classification tasks, these scores are not the likelihood probabilities. If that was the case, we would use Softmax to map logits scores into the probability domain and select the class with the highest probability. However, in our case we have multiple values in our hands, which needs to be predicted as conflict and positive. In the most basic approach, if we had to select half of them conflict and half positive, the straightforward operation would be finding the median value and split the instances choosing that value as the threshold following the prediction. The higher scores would be predicted as conflicts and the others would be positives. Since we do not force the model to select a specific number of conflict samples, we should determine a threshold level to split the instances. If we would like to punish missing conflict cases more severe, our threshold would be lower in order not to miss many conflict instances, but this would increase false negatives (if conflict counts as the negative). In another case, while trying to decrease the false negatives, this time false positives would increase. In this kind of cases where we do not specify our precedences, it is pretty common to use AUC (Area Under the Curve) metric instead of accuracy.

Not being able to use accuracy directly, is not the only reason why we use AUC while evaluating our performance. In [59], authors assert many different reasons why AUC could be a better evaluation measure than accuracy while comparing classifiers.

The "curve" in AUC is Receiver Operating Characteristic (ROC) curve, which is simply the plot of the True Positive Rate (TPR) to False Positive Rate (FPR). The point (0, 0) is where we do not predict any positives, and (1, 1) is where we predict positive for every sample. A successful model would have a curve closer to the upper left corner, which increases the area under the curve, so does our AUC score. It means the model outputs True Positives without giving so many False Positives, as desired.

It should be noted that our proposed model is optimized according to the AUC measure, where the rest of the experiments are optimized with accuracy. We cannot directly compare those two different measures numerically, but as stated in [60], AUC is a summary measure that essentially averages the accuracy across the spectrum of testing values. Also, 0.5 and 1 values correspond to the same results, as "random choice" and "perfect classification" extremes, respectively. Therefore, assuming the values reported as AUC or accuracy would be highly correlated could be accurate. However, in order to be able to compare the results fairly and be as precise as possible, we provide proposed model results both in AUC and accuracy measures. The accuracy values of the proposed model are obtained using Equal Error Rate (EER). The False Positive Rate (FPR) and False Negative Rate (FNR) versus threshold plots intersect at one point, since one is monotonically increasing, where the other is decreasing. For finding the intersection, we interpolate the FPR and FNR values. When these rates are equal (intersection), this common value is referred to as the Equal Error Rate, which corresponds to the optimal threshold found with this method, since we penalize missing positive and conflict cases equally. Then, we calculate the accuracy values for our proposed models by subtracting this value from 1, which are compared with the baseline scores in Sections 3.3.3 and 3.3.4.

In tables where we do not share fold specific results, we report the overall results from the whole dataset. Since our splits has different number of instances in them, we do not directly take the average of folds while reporting the final results, but they are very similar with the averages since the instance count difference is very small. The first 4 folds contain 18 instances and the rest contains 19 instances per folds, so, a weighted average would work fine.

3.3 Results

In this section, we share experimental results on different analyses such as the outcomes and techniques of processing stand alone individuals or couples in pairs and the fusion techniques when pairs are used, the model reliability according to different fold results, comparison of proposed method with baseline results we provide along with the human evaluation studies.

3.3.1 Individual Subject versus Pair Analysis

The proposed network predicts a score for each of the partitions. This score indicates how likely the video partition is from a conflict session. That is, as the scores increase, the probability that we predict this video sequence as conflict also increases. In Figure 3.3, we present an example score distribution in order to show segment scores in more detail. It can be seen that the positive session is mostly differentiated from the conflict session. When we check the maximum score of each session, we see that conflict has a higher one, which indicates the classification is successful. After generating the ROC curve, if we draw the separating threshold somewhere between [0.48, 0.50], then both of these values would be counted as true positives (100% success). When the threshold is above the conflict maximum, true positives would be one less and when it is below the positive maximum value, there will be one more false positive (50% success in each case). If the model wrongly predicts the conflict-positive separation, then the corresponding success values would be 0% and 50%, respectively.

The scores in Figure 3.3 are assigned to each video segment in the bag, which indicates the segment's conflict level. We look for the maximum valued segment in order to predict the label for the whole video, as explained in Section 2.3. A simple but effective approach to gather two views showing both faces of a session is, not to predict video-level labels, but instead predicting session-level labels, by taking the maximum of the whole session. Previously, our bag was involving the instances from one video where we would look at a single participant. Now, the experiments in the third and fourth rows in Table 3.2, we train the same way, but while making our predictions, we look at the maximum of all instances in both person's bags. Here, our motivation is that, in conflict sessions, while a participant may stay calm during the video and does not reveal easily that the session label is conflict, the other might give stronger clues that it is indeed a conflict session. So, we simply predict "conflict" for the session, if one of the participants is predicted as "conflict".

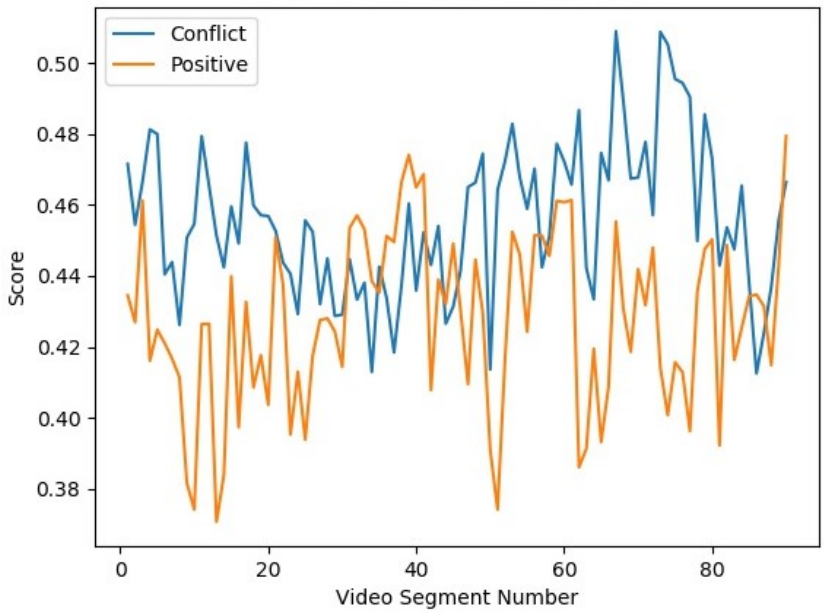


Figure 3.3: Example positive experience and conflict session scores, where the video is divided into 90 segments

Model in second row of Table 3.2 applies the early fusion, meaning that concatenating the 512D feature vectors before feeding them to the proposed model, by making 1024D vectors. Some of the models that apply late fusion are not optimized to this technique, meaning that the training procedure is the same as the individual subject case, but while evaluating the validation and test sets, we look at the scores as explained in the previous paragraph. The models optimized with this technique have their training sets evaluated and gradients calculated accordingly. Surprisingly, early fusion technique and Max of Session optimized models perform poorly in test sets, whereas their validation scores are really close to those of our best model. The early fusion might not represent the participants as well as the original features, and the late fusion with optimization might be harmful to the backpropagation process. However, the late fusion without the optimization reaches 74.48% in AUC, which is the best result we obtain. Although taking the maximum of both sessions might seem simple, it is very effective benefiting from the connections between each participant and it increases the best individual MIL result by 11.65% with respect to the individual test AUC. From now on in this thesis, this model would be referred as the "proposed model".

Table 3.2: Results of the MIL Model

Subject of Interest	**Fusion**	**Technique**	**Validation AUC**	**Test AUC**
Individual	-	Max of Bag	67.12	66.71
Pair	Early	Max of Bag	73.77	56.72
Pair	Late - With optim.	Max of Session	73.35	56.04
Pair	Late - No optim.	Max of Session	74.02	**74.54**

In order to visualize the ROC curves and to understand which AUC values we obtain, the best and worst performing folds' ROC curves and corresponding AUC values from the proposed model are shown in Figure 3.4.

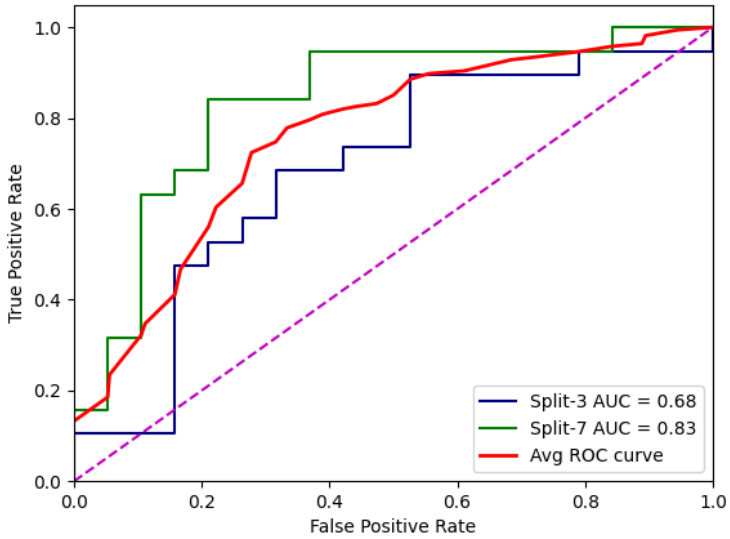


Figure 3.4: ROC curves and corresponding AUC results generated from the proposed model results in Table 3.2. Results of the split that yields the lowest AUC score (3th), the highest AUC score (7th) and average ROC curve of all splits are shown with blue, green and red, respectively

We could have used ResNet-3D and proposed architecture without using the extracted representations, as an end-to-end architecture with finetuning the parameters of ResNet-3D to enhance our results. However, even the feature extraction process is extremely heavy in terms of time and trying to finetune the network to extract features more specified for our task would be a real burden for our network. Considering that we have so many folds and hyperparameter sets, and the fact that these are all independent variables and increase the experiment count exponentially, we choose using the ResNet-3D architecture as frozen and separated the feature extraction and training stages.

3.3.2 Effect of Behavioral Data Differences on Model Reliability

Table 3.3: Effect of Behavioral Data Differences on Model Reliability

Fold Number	Validation AUC	Test AUC	Relative Difference %
0	m0 - 78.40	m8 - 77.78	0.77
1	m1 - 71.30	m0 - 71.60	0.42
2	m2 - 75.62	m1 - 75.31	0.40
3	m3 - 72.22	m2 - 71.91	0.42
4	m4 - 68.42	m3 - 68.42	0.00
5	m5 - 71.47	m4 - 70.36	1.54
6	m6 - 73.13	m5 - 73.96	1.23
7	m7 - 72.58	m6 - 77.56	6.89
8	m8 - 83.10	m7 - 83.38	0.36
Combined	74.02	74.54	1.34

In Table 3.3, we can observe the best validation and test results for each *fold* (not split). One might think that each row belongs to a model and the test results are coming from the same model that gives the corresponding validation results, but it would be wrong. In order to prevent this possible confusion, we write the model numbers next to the AUC scores, where mn indicates the n^{th} model trained with split-n. It can be observed that regardless the split, the obtained AUC scores are very similar for each specific fold. The absolute difference percentage showing the proximity of test and validation scores of a fold, obtained from different splits, is calculated by dividing the absolute difference by the validation results. As referenced in Section 3.2.2 as the "following pages", it can be seen that the absolute difference percentage between test and validation results is 1.34% when combined, which is very low. Therefore, we can deduce that although the model neither underfits nor overfits the data, independent from whichever folds form the train set, it fits to a very close state. The reason why there are differences in AUC scores is not only related with the model performance, but also about the randomly generated folds' prediction difficulty differences.

3.3.3 Comparison to other Methods

Since there is no previous work done on this dataset, we implement several other methods to provide baselines for our challenge. These baselines include end-to-end spatio-temporal modelings, temporal models with frozen features and support vector classifiers.

We use several spatio-temporal architectures, where we can simultaneously create connections between pixels and their changes over time. There are three main networks with variations we use:

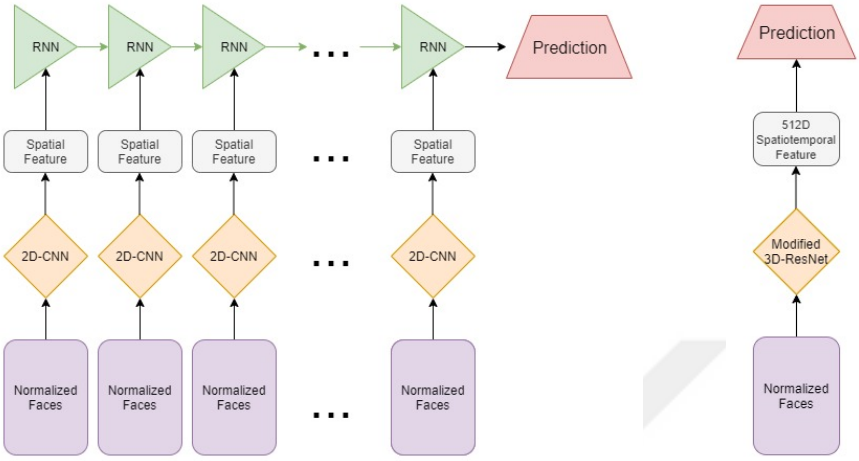
- CNN - RNN architecture with vanilla RNN, GRU and LSTM networks as in Figure 3.5 (a) fed with warped faces and raw frames, in order to observe whether the face normalization is necessary or not for this architecture.
- ResNet-3D [56] fed with warped faces, having single or multiple passes through the network
- Bidirectional LSTM (BiLSTM) network fed with frozen ResNet-3D features

CNN-RNN architecture is shown in Figure 3.5 (a). We use AlexNet [61] as our CNN architecture, removing its last layer. Variations in the CNN-RNN experiments include the weights being trained from scratch, as well as using the weights pretrained on ImageNet, as indicated in [61]. The pretrained weights are finetuned or used frozen in different experiments. Vanilla RNN, GRU and LSTM are used as the recurrent network and in order to get video-level predictions, we use many-to-one strategy. In a similar method, we use BiLSTMs, with frozen ResNet-3D features being the input, instead of the images used in CNN-RNNs.

Single-pass and multi-pass architectures adapted from ResNet-3D are shown in Figure 3.5 (b) and (c), respectively, where modified ResNet-3D is the architecture shown in Figure 2.2. The last fully connected layer is dropped out since we do not want to use that overly specified weights in our network, and the penultimate 512D feature is being used. In single-pass architecture, we directly finetune the

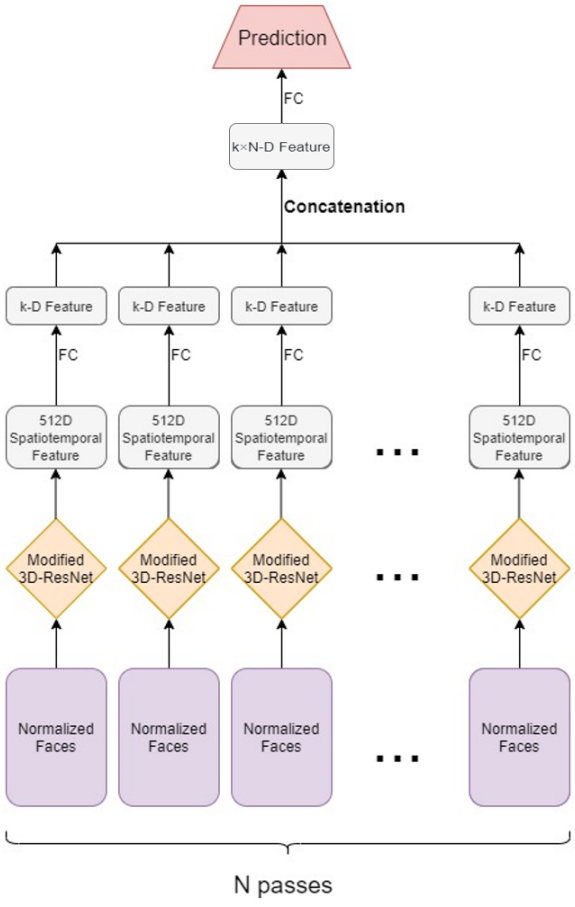
pretrained model giving all normalized faces from a single subject of interest. In multi-pass architecture, we introduce two hyperparameters; the number of "passes" (N) and the intermediate embedding size (k). As we can see in Figure 3.5 (b), this architecture is consisting of multiple single-pass ResNet-3D's, with a slight variation of concatenation of the output vectors from video segments. Our goal is to divide the video in smaller segments and get specific feature vectors from each of them. All these features are concatenated in the early fusion layer and fed to a Fully Connected (FC) layer again. This way, the model can learn the connections between different video segments and their corresponding feature vectors. The prediction layer is 1D output layer in each architecture, where we use Binary Cross Entropy loss to conduct binary classification.

During the SVM experiments, we preprocess the features in 3 different ways. For fixed partitions methods with partition count 120 and ResNet-3D features with 512 dimensions, one video instance would have a shape of (120, 512). In flatten feature feeding type, which is the best performing, hence reported technique, we directly obtain (1, 120×512) shaped features to be fed to SVM's. Timewise and dimensionwise are the other techniques which are self-explanatory; in timewise processing we have (1, 512) shaped features, whereas in dimensionwise processing case the shape becomes (1, 120). We also concatenate a set of the minimum, maximum and standard deviation features while conducting hyperparameter search using each technique. For example, if we are using all 4 mathematical operations in timewise processing, we would have (1, 4×512) features in our hands.



(a)

(b)



(c)

Figure 3.5: CNN-RNN (a), ResNet-3D Single-pass (b) and ResNet-3D Multi-pass (c) architectures

Table 3.4: Accuracy Comparison to Different Models

Architecture	**Input**	**Subject of Interest**	**Test Accuracy**
Proposed	Frozen ResNet-3D FPa	Pair	**71.01**
		Individual	63.02
SVM		Individual	67.22
		Pair	63.25
BiLSTM	Frozen ResNet-3D FPe	Individual	56.44
ResNet-3D MP	Norm. Frames		53.89
ResNet-3D SP			51.50
LST-Net	AU Intensities		51.13
CNN-LSTM	Norm. Frames / Raw Frames		50.00

FPe: Fixed Period, FPa: Fixed Partitions, AU: Action Units,
MP: Multi-pass, SP: Single-pass

We report the proposed model and baseline test accuracies for both pair and individual subject of interests in Table 3.4. The reported values are the best results within their method, that is, none of the other variations of these architectures we try can get better scores than these ones. In Table 3.4, the proposed model result is given in accuracy rather than AUC value reported in Table 3.2, in order to be able to compare with the other methods, whose scores are optimized for higher accuracies. Although we state that the AUC metric is nearly comparable with accuracy results, due to unavailability of selecting a perfect threshold for separation, our proposed model gives 71.01% in accuracy, with a drop of 4.73% with respect to its AUC score. One reason why proposed model's accuracy is lower than its AUC value is that while we optimize our proposed models and selecting the best performing model, we try to maximize the AUC value instead of accuracy. If the whole process is designed to optimize accuracy the chances are high that we obtain a better result. However, we believe that reporting AUC results is much better for our proposed model, since we can adjust the thresholds according to our desire of favoring true positive rates over true negative rates or vice versa. In any way, although the results drop a bit due to this imperfection,

our proposed model still outperforms any of the other methods.

When we check the other methods and the differences between their variations, we see that the SVM experiment that takes subjects as pairs has a lower test accuracy than its individual counterpart. The concatenation doubles the dimension, which might have resulted in a worse model due to the curse of dimensionality when using ResNet-3D features. The pair model might also experience overfitting, compared to its individual counterpart. In any case, SVM baseline results show that we can obtain valuable information from facial cues about the session's state.

One can observe that the individual subject SVM accuracy is just 0.51% better than the individual subject proposed architecture AUC result (shown in Table 3.2). With the drop in accuracies with respect to AUC values for proposed individual model, this difference increases to 6.25%. Although reported SVM results are better for individuals, while it cannot improve the results any better by using pairs as subject, our proposed model benefits the availability of pairs. Only concatenating the features in SVM seems to be not enough for improvements, whereas the proposed model is more successful in our main task, which is classifying *dyadic* interactions. Here, the utilization of multiple instance learning outperforms the direct concatenation and benefit from using multiple views of pairs more efficiently, since proposed model using pairs improved the model using individuals by 12.68% with respect to the individual accuracy.

While using CNN-RNN structure, the main problem we face was the vanishing gradients. Since the videos are so long, direct classification using the whole video mostly underfit the data. We observe that the gradients die even in the first epoch and all the instances, regardless of their labels, are predicted as the same score. The scores are changing between the epochs, but still the same for all instances. In Table 3.4, we can observe that the CNN-RNN models (either used with Vanilla RNN, GRU or LSTM) cannot learn the weights necessary for this binary classification, as they all have 50% test accuracies. Intuitively, we know that random choice in binary classification would give 50% accuracy, which would not be enough. We try raw frames instead of normalized faces as inputs, since

we would we the scene at the same time, like body and head movements. Using several different hyperparameter settings, activating and deactivating bias are among the experiments we try, but none of these alterations change the results. The nature of this architecture might not be matching with the needs of our task.

LST-Net [62], which is a time-series model, gives slight improvement when fed only with action units, compared to adding head pose and gaze features, which does not support the action units by resulting in 50% accuracy. The reason of this might be that action units and gaze-head pose are not directly correlated and while inferring the positivity of a session from head pose and gaze, those features probably introduce irrelevant information that would cancel out the slight success of action units. Although action units can perform well in many cases, our proposed model got way better results in this classification task.

We can observe slight improvements in ResNet-3D model, especially when used with multi-pass technique. Here, the N value which denotes the number of passes, is selected as 90 and the embedding size is 64. Although the results are still way lower than those obtained from SVM's, the fact that multi-pass technique gave better results than single-pass is promising for the MIL method that we use. The main weakness of multi-pass technique compared to our proposed model is that it uses neural networks (MLP) to predict the label, where we cannot know the exact moment of interest. By the concatenation of intermediate embeddings, the model gives importance to the order in which it sees the input, due to the nature of neural networks and weight matrices. However, in our proposed model, by using the bags, we do not concentrate on the moment of actions, we rather want to obtain the information whether is exists or not throughout the video. This flexibility in time, while also processing the shorter-term temporal information through the input features is probably one of the reasons why our proposed model outperforms the multi-pass ResNet-3D model.

3.3.4 Human Evaluation

As this dataset is used for the first time in the literature, there is not any state of the art (SOTA) results for the same task. We provide the baseline and our own SOTA in Section 3.3.3, but we also think that providing some numerical results obtained in a different way would support the discussion of our success of classification. Since there is no other computational competitor, it might be beneficial to observe how real humans succeed in a task very similar to ours, in this dataset. To this end, we have found volunteers to classify videos as conflict or positive, the same way our model does, following a specific procedure. Our initial motivation was not surpassing human prediction, even getting close to it and having comparable scores with humans would be very meaningful.

Showing all the videos to each evaluator would take so much time, so we decided to select a representative subset of videos to show to the human evaluators. While selecting those videos, in order to avoid easiest and hardest samples in the dataset, we eliminate the 3 folds with the highest scores and the 3 folds with the lowest scores in validation results of our proposed model. The 3 folds with medium successes are seen to be the fold numbers 3, 6 and 7 as observable in Table 3.3.

We have shown all of the instances in these 3 folds, to the completely new participants who are involved in human evaluation part. That is, these are not the 10 professional annotators described in Section 3.1, who labeled the data. There were in total 6 evaluators, 3 male and 3 female, aged in between 18-58 and these people are not trained for this kind of prediction, like the annotators. However, the set of participants included psychologists and medical doctors, which might affect their predictions in a supportive way because of their professions.

The procedure of human video evaluation is as follows: The video clip to be shown is selected randomly among the predefined 3 folds. At first, a video that involves only one person is selected (either a male or a female). Being female or male, coming from a positive or a conflict video is also selected in random order, in order to prevent any possible side effects stemming from the order we show the

videos. We use the video clips as they are recorded, that is, no further processing like face-cropping is applied to the videos. So, it should be noted that human evaluators are able to see the body language and the environment as they see the full frame, whereas in most of our models we only work with the normalized faces and their representations. Although we have voices in the videos, in order to force evaluators to concentrate on the visual effects, we mute the videos. Otherwise, as the evaluators are all Turkish speakers and in the videos people are speaking in Turkish, they would listen to the meaning of the speech and probably get as high results as 100%. Since we do not apply any NLP or audio models in this thesis and our concentration is on computer vision, we gathered the human results only with the visual aspects.

Since the recordings are approximately 10 minutes and there are 56 couples that we are showing to the evaluators, we cannot ask them to watch the full videos and evaluate them accordingly. Therefore, we use a technique that is effectively utilized in psychology domain, which is called "thin slices" [22]. In their research, Ambady and Rosenthal showed that while having judgments assessing nonverbal behavior of a person, a complete stranger can reach accuracies as high as the people who had substantial interactions with that specific person in real life, just by watching video clips from 2 seconds to 10 seconds (thin slice). Furthermore, watching 3×10 seconds of videos does not significantly increase the accuracy, in comparison with 3×2 seconds. In 2017, Gunaydin et al. [63] uses this technique to show the videos to 8 independent coders, without any audio again, and ask them to evaluate the warmth, partner interaction and engagement of a person. Hence, we show the videos to our independent evaluators in three sessions, all of which are at least 2 seconds (in total at least 6 seconds), but in confusing cases, the evaluator can prolong this duration as much as he or she wants. Since the conversations are unsupervised, that is, the couples are talking in whatever direction they would like to continue, we pick the starting times of those 2 or more seconds randomly, but avoided the first and last 1 minute to ensure we are not interrupted by the researcher which appears in the end and the beginning of the videos. With this random selection, at times, evaluator watches the person listening to the other one, or actively speaking; which is what our model sees

in general. At this point it is worth reminding that our models see the video partitions of 3-8,5 seconds approximately, so we can relate this human prediction of seeing 3 partitions consisting of at least 2 seconds to our model selection where we process many more partitions, but with similar lengths.

Our models are not evaluating the comparison of "which one of the two sessions of a couple is *more* positive", but rather they are evaluating the videos individually on being positive or conflict. However, if we have shown both of the sessions to a human evaluator, most probably he or she would immediately compare the two videos in his or her mind. To avoid this issue, if the evaluator sees a dyad's conflict session, we did not show him or her the positive session of the same dyad. Therefore, every session of the couples are evaluated by half of the independent evaluators.

The evaluators start investigating a dyad by watching either the male or the female participant. After making all the predictions watching these single participants, we show, again in a random order, the other participant of the same couples, same session and same minutes along with the initial participant, and make them re-predict for these sessions where they are evaluating both parties in pairs this time. We also run our models with individuals or pairs being the subject of interests. To study the effect of being able to see both participants at the same time and compare it with our results, this re-prediction can be very beneficial.

Table 3.5: Human Evaluation Accuracies

	Accuracy			
Subject of Interest	**Mean**	**Min**	**Max**	**Std**
Individual	65.18	60.71	78.57	6.84
Pair	64.58	55.36	69.64	5.35

Looking at the results in Table 3.5, we observe that our independent evaluators' success rate decreases slightly when they saw both of the couples. While showing the videos, it was noted that they have spent less time investigating the videos

that they saw in the later stages. In the first videos, they looked at the video longer than 6 seconds so many times, but this number get less and less in the last videos. So, this might be one reason why investigating both participants at the same time causes a drop in their prediction accuracies.

Table 3.6: Human Evaluation Prediction Transitions Between Individual and Pair Sessions

Transition %	**To Correct**	**To Incorrect**
From Correct	71.69	28.31
From Incorrect	51.28	48.72

In Table 3.6, we list the percentages of transitions between individual and pair sessions of accuracies. For example, if the evaluator correctly predicted the session after watching one individual's video and changed his or her idea after watching both of the couples' videos in pair, it is counted to be in the top-right cell. Here, we can see that after seeing both participants, many evaluators predicted differently, both correctly and incorrectly at the same time. The correct-to-incorrect transition is a bit higher than our expectations, which might imply that humans may not be so successful concentrating on two faces at the same time and they can miss the clues that are caught by them beforehand. Since the correct predictions are more than the incorrect predictions, 28.31% change in correct ones had more effect than the 51.28% change in the incorrect ones, resulting in a poorer performance in watching pairs.

Table 3.7: Comparison of Proposed Models and Human Evaluation

Evaluator	**Subject of Interest**	**Metric**	**Test Result**
Human	Individual	Accuracy	65.18
	Pair		64.58
Proposed Model	Individual	AUC	66.00
		Accuracy	63.51
	Pair	AUC	**74.52**
		Accuracy	**69.64**

Overall, we observe that individual human results have slightly better accuracy values than proposed individual accuracies, even though the AUC result is higher for the proposed method, which might suggest when the optimization for AUC value and not accuracy imperfection described in Section 3.3.3 issues are overcome, the individual accuracies can also beat humans. Switching to pairs, our model clearly outperform human performance when we use both of the couples. Our best model probably created better connections between each participant and benefiting from the action-reaction mechanism of these dyadic communications, it was able to enhance its results. It should be noted that, this table shows the results only for folds 3, 6 and 7. That is the reason of the change in results. Proposed model AUC result does not change much, which is expected since we select the medium performing 3 folds.

3.4 Failure Cases and Discussion

Our proposed model produces arguably good results for the classification of romantic relationship interactions. The Romantic Relationship Dataset used in this study, includes equal number of positive experience and conflict samples. In other words, there is no class imbalance in the dataset. In a real life scenario, on the other hand, this may not be the case. It is possible to observe much more positive experience samples than the conflict ones. In case of training with such a dataset, penalty for conflict misclassification should be increased. Consequently, modified models would be able to produce more conflict predictions, in order to increase the prediction reliability for conflict cases. If we train a basic statistical model on a balanced dataset, and test it on an imbalanced one where we have much more negative samples, then the predictions for many negative samples would tend to be positive. Due to such issues, in imbalanced scenarios, precision/recall analysis would be a better choice for optimization.

To comprehend the visual patterns that shape the classification results, we have analyzed the instances (MIL instances) with highest scores since the maximum scored instance determines the final prediction in our model. In many

misprediction cases of positive experience samples, we observe that the participant stays still, and displays very slight or no facial expressions. In these instances of the videos, the target participant is actually listening to the other participant, but the stable appearance of his/her face convinces the model that this is not a positive instance. In the opposite case, we observe that the instances where the participants slightly smile/laugh or display more expressions, result in positive predictions for the conflict sessions. These findings can be expected due to the nature of MIL model that focuses only on the instance with the highest score, where the corresponding instance may not effectively represent the whole session. The comparably low human evaluation results support these findings as humans can also be deceived when they observe such facial behavior in the analyzed video segments. Sample frames cannot be presented due to copyright issues.

A smile, laughter or similar positive expressions that appear for a short period in the beginning or at the end of a video segment coming from a conflict case may decrease its predicted conflict score, deceiving the model that this instance may be coming from a positive instance.By dividing the video into segments, we may increase the individual effects of those expressions on the prediction of those segments, since the period we investigate is much shorter now and their influence is hence increased. However, since we only use the maximum scored instance in Multiple Instance Learning bags, these deceptions would have a minor effect. Even if we observe such deceiving instances and mispredict the conflict instance as positive, since the nature of Romantic Relationship Dataset does not generally include only one conflicting moment throughout the session, there would probably be another instance that our model would catch as a conflicting segment. This would totally eliminate the misprediction effect of that instance. For example in Figure 3.3, if we had missed the maximum valued instance due to this kind of an error, still we would predict the whole video as conflict since there are some other highly rated segments in that video.

As described earlier, based on the use of multiple instance learning, we select the maximum scored instance as the representative segment of the whole video. Consequently, other video segments do not affect the learning process. Yet, due to several reasons such as noise and confusing behavior, relying solely on the

maximum scored instance may be problematic in some samples. Such issues may be reduced by selecting not one, but more high scored instances, and using their average as the representative score. On the other hand, by doing so we would contradict with our assumption that the conflict behavior may happen very occasionally, even only in one segment of a conflict case sample. Consequently, we may face other issues while minimizing the effect of noise and confusing patterns with the use of average scores.

While the proposed approach may be used for several applications in human behavior analysis, reaching high-stake decisions solely based on automated analysis would not be acceptable from an ethical point of view. In addition, the accuracy of an automated system on a sample population may not represent its actual reliability. Therefore, all ethical aspects should be carefully considered before employing such intelligent systems in real life scenarios. Furthermore, researchers have to be cautious with their claims on the potential use cases of their proposed models.

Chapter 4

Conclusion

In this thesis, we have designed and implemented a dyadic interaction classification framework that uses the videos of romantic couples who are talking on topics they have conflicts on in one setting, and positive experiences they share in the other. To this end, we model facial cues to classify these sessions, where we combine deep multiple instance learning with deep metric learning. The instances in the bags formed for the multiple instance learning, are selected as the feature vectors that we extract from the video segments using ResNet-3D pretrained model. In this model, our aim is to increase the vectorial distance between the visual representations of video segments belonging to different classes. The maximum value in the bag of video segments would represent the overall conflict level of the video, and it is expected to be higher for many of the conflict videos if the separation is successful.

We have evaluated different variations of the proposed model, such as using the facial expressions of individuals or those of the pairs jointly as the subjects of interest. For analyzing pairs jointly, we employ early and late fusion where we optimize the network for the fused representation or directly fuse the prediction scores. The best approach has been found to be the not-optimized late fusion with pair inputs, reaching an AUC of 74.54%. To the best of our knowledge, this is the first attempt to the automated classification of dyadic interactions of

romantic couples. Therefore, there is no available competitor methods for this task. Hence, we have implemented several baseline models such as employing 3D Residual Networks, Support Vector Machines, Long Short-Term Memory and Convolutional Neural Networks, and provided comparisons. In our experiments, the best performing baseline reaches an accuracy of 67.22%, where that of the proposed approach is 71.01%.

When analyzing the baseline results, we observe that SVM has reached the most accurate results, outperforming the spatio-temporal baseline models. This finding may suggest that the spatio-temporal models have issues to form strong temporal connections, possibly due to the very long duration of the input videos (e.g. about 10 minutes).

After investigating the fold specific results, we observe that the best performing models have performed similarly on average for the same folds. Yet, there is a clear difference in the prediction accuracy for different folds. This can be explained by the fact that the facial behavior differ significantly between different subjects during dyadic interactions.

We have also conducted a human study, where six participants conducts the classification task and obtains an accuracy of 65.18%. Interestingly, humans could not perform better when they analyze the videos of both of the couples, compared to the analysis of one individual's video. Joint analysis of subjects even decreases human accuracy by 0.93%. This may be due to the fact that the video short segments analyzed by participants are randomly chosen. On the other hand, the use of videos of both of couples in the automated analysis, improves the classification performance by 9.65%.

We can conclude that, using facial expressions can indeed be used to infer the state of a dyadic interaction and our proposed model is performing in a reliable manner, providing even better results than humans. It is important to note that in such a complex and abstract domain like human behavior analysis, human predictions mostly rely on intuitions and experiences that are obtained through life and this is arguably a harder problem for a computational model to solve just

based on limited amount of training data.

As a future work, the task can be approached in a multimodal manner, where different modalities such as voice and speech (language) can provide additional information. In addition, in case of an accurate tracking, side view of body pose can also be employed in the analysis. Furthermore, increasing the number of participants in the human evaluation and providing them longer sequences of couple's behavior may also yield a better human accuracy.

Bibliography

[1] T. Ledermann, G. Bodenmann, S. Gagliardi, L. Charvoz, S. Verardi, J. Rossier, A. Bertoni, and R. Iafrate, "Psychometrics of the dyadic coping inventory in three language groups," *Swiss Journal of Psychology*, 2010.

[2] P. Thomas and W. Fraser, "Linguistics, human communication and psychiatry," *The British Journal of Psychiatry*, vol. 165, no. 5, pp. 585–592, 1994.

[3] S. Petrocchi, P. Iannello, F. Lecciso, A. Levante, A. Antonietti, and P. Schulz, "Interpersonal trust in doctor-patient relation: Evidence from dyadic analysis and association with quality of dyadic communication," *Social Science & Medicine*, vol. 235, p. 112391, 2019.

[4] N. Jatupaiboon, S. Pan-Ngum, and P. Israsena, "Subject-dependent and subject-independent emotion classification using unimodal and multimodal physiological signals," *Journal of Medical Imaging and Health Informatics*, vol. 5, no. 5, pp. 1020–1027, 2015.

[5] S. T. Kinney and R. T. Watson, "The effect of medium and task on dyadic communication," 1992.

[6] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3460–3469, 2015.

[7] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 189–203, 2016.

[8] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, 2018.

[9] S. Phan, D.-D. Le, and S. Satoh, "Multimedia event detection using event-driven multiple instance learning," in *ACM International Conference on Multimedia*, pp. 1255–1258, 2015.

[10] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE Reviews in Biomedical Engineering*, vol. 10, pp. 213–234, 2017.

[11] K. Wu, C. Liu, and R. A. Calvo, "Automatic nonverbal mimicry detection and analysis in medical video consultations," *International Journal of Human–Computer Interaction*, vol. 36, no. 14, pp. 1379–1392, 2020.

[12] E. O. Cheung, E. B. Slotter, and W. L. Gardner, "Are you feeling what i'm feeling? the role of facial mimicry in facilitating reconnection following social exclusion," *Motivation and Emotion*, vol. 39, no. 4, pp. 613–630, 2015.

[13] H. Drimalla, N. Landwehr, U. Hess, and I. Dziobek, "From face to face: the contribution of facial mimicry to cognitive and emotional empathy," *Cognition and Emotion*, 2019.

[14] G. Varni, I. Hupont, C. Clavel, and M. Chetouani, "Computational study of primitive emotional contagion in dyadic interactions," *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 258–271, 2017.

[15] T. L. Chartrand and J. A. Bargh, "The chameleon effect: the perception–behavior link and social interaction.," *Journal of Personality and Social Psychology*, vol. 76, no. 6, p. 893, 1999.

[16] M. LaFrance, "Postural mirroring and intergroup relations," *Personality and Social Psychology Bulletin*, vol. 11, no. 2, pp. 207–217, 1985.

[17] E. Hatfield, J. T. Cacioppo, and R. L. Rapson, "Emotional contagion," *Current Directions in Psychological Science*, vol. 2, no. 3, pp. 96–100, 1993.

[18] M. Iacoboni, *Mirroring people: The new science of how we connect with others*. Farrar, Straus and Giroux, 2009.

[19] M. LaFrance and M. Broadbent, "Group rapport: Posture sharing as a nonverbal indicator," *Group & Organization Studies*, vol. 1, no. 3, pp. 328–333, 1976.

[20] F. J. Bernieri, "Coordinated movement and rapport in teacher-student interactions," *Journal of Nonverbal Behavior*, vol. 12, no. 2, pp. 120–138, 1988.

[21] J. Zhou, "The effects of reciprocal imitation on teacher–student relationships and student learning outcomes," *Mind, Brain, and Education*, vol. 6, no. 2, pp. 66–73, 2012.

[22] N. Ambady and R. Rosenthal, "Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness.," *Journal of Personality and Social Psychology*, vol. 64, no. 3, p. 431, 1993.

[23] T. R. Lynch, *Radically open dialectical behavior therapy: Theory and practice for treating disorders of overcontrol*. New Harbinger Publications, 2018.

[24] K. Gilbert, K. Hall, and R. T. Codd, "Radically open dialectical behavior therapy: Social signaling, transdiagnostic utility and current evidence," *Psychology Research and Behavior Management*, vol. 13, p. 19, 2020.

[25] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 69–87, 2011.

[26] A. A. Salah, M. Pantic, and A. Vinciarelli, "Recent developments in social signal processing," in *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 380–385, IEEE, 2011.

[27] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[28] A. Stergiou and R. Poppe, "Analyzing human–human interactions: A survey," *Computer Vision and Image Understanding*, vol. 188, p. 102799, 2019.

[29] P. Chopade, S. M. Khan, D. Edwards, and A. von Davier, "Machine learning for efficient assessment and prediction of human performance in collaborative learning environments," in *IEEE International Symposium on Technologies for Homeland Security*, pp. 1–6, IEEE, 2018.

[30] M. Doyran, A. Schimmel, P. Baki, K. Ergin, B. Türkmen, A. A. Salah, S. C. Bakkes, H. Kaya, R. Poppe, and A. A. Salah, "Mumbai: Multi-person, multimodal board game affect and interaction analysis dataset," *Journal on Multimodal User Interfaces*, pp. 1–19, 2021.

[31] S. Halfon, M. Doyran, B. Türkmen, E. A. Oktay, and A. A. Salah, "Multimodal affect analysis of psychodynamic play therapy," *Psychotherapy Research*, vol. 31, no. 3, pp. 313–328, 2021.

[32] Y. Zhang, J. Olenick, C.-H. Chang, S. W. Kozlowski, and H. Hung, "TeamSense: Assessing personal affect and group cohesion in small teams through dyadic interaction and behavior analysis with wearable sensors," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–22, 2018.

[33] J. R. Terven, B. Raducanu, M. E. Meza-de Luna, and J. Salas, "Head-gestures mirroring detection in dyadic social interactions with computer vision-based wearable devices," *Neurocomputing*, vol. 175, pp. 866–876, 2016.

[34] G. Calabrò, A. Bizzego, S. Cainelli, C. Furlanello, and P. Venuti, "M-MS: A multi-Modal synchrony dataset to explore dyadic interaction in ASD," in *Progresses in Artificial Intelligence and Neural Systems*, pp. 543–553, Springer, 2021.

[35] B. Xiao, P. Georgiou, B. Baucom, and S. S. Narayanan, "Head motion modeling for human behavior analysis in dyadic interaction," *IEEE Transactions on Multimedia*, vol. 17, no. 7, pp. 1107–1119, 2015.

[36] J. L. Hagad, R. Legaspi, M. Numao, and M. Suarez, "Predicting levels of rapport in dyadic interactions through automatic detection of posture and posture congruence," in *IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing*, pp. 613–616, IEEE, 2011.

[37] A. S. Won, J. N. Bailenson, and J. H. Janssen, "Automatic detection of nonverbal behavior predicts learning in dyadic interactions," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 112–125, 2014.

[38] S. N. Chakravarthula, B. Baucom, and P. Georgiou, "Modeling interpersonal influence of verbal behavior in couples therapy dyadic interactions," *arXiv preprint arXiv:1805.09436*, 2018.

[39] F. Ramseyer and W. Tschacher, "Synchrony: A core concept for a constructivist approach to psychotherapy," *Constructivism in the Human Sciences*, vol. 11, no. 1-2, pp. 150–171, 2006.

[40] W. Tschacher, G. M. Rees, and F. Ramseyer, "Nonverbal synchrony and affect in dyadic interactions," *Frontiers in Psychology*, vol. 5, p. 1323, 2014.

[41] A. Cerekovic, O. Aran, and D. Gatica-Perez, "How do you like your virtual agent?: Human-agent interaction experience through nonverbal features and personality traits," in *International Workshop on Human Behavior Understanding*, pp. 1–15, Springer, 2014.

[42] Z. Hammal, J. F. Cohn, and D. S. Messinger, "Head movement dynamics during play and perturbed mother-infant interaction," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 361–370, 2015.

[43] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[44] J. M. Girard, W.-S. Chu, L. A. Jeni, and J. F. Cohn, "Sayette group formation task (gft) spontaneous facial expression database," in *IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 581–588, IEEE, 2017.

[45] N. Wang and J. Gratch, "Rapport and facial expression," in *International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–6, IEEE, 2009.

[46] Y. Huang and S. M. Khan, "Dyadgan: Generating facial expressions in dyadic interactions," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 11–18, 2017.

[47] Manual of Emotient's Facet SDK., "iMotions Inc.," 2013.

[48] B. Nojavanasghari, Y. Huang, and S. Khan, "Interactive generative adversarial networks for facial expression generation in dyadic interactions," *arXiv preprint arXiv:1801.09092*, 2018.

[49] D. Giritlioğlu, B. Mandira, S. F. Yilmaz, C. U. Ertenli, B. F. Akgür, M. Kınıklıoğlu, A. G. Kurt, E. Mutlu, Ş. C. Gürel, and H. Dibeklioğlu, "Multimodal analysis of personality traits on videos of self-presentation and induced behavior," *Journal on Multimodal User Interfaces*, pp. 1–22, 2020.

[50] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019.

[51] H. J. Escalante, H. Kaya, A. A. Salah, S. Escalera, Y. Güçlütürk, U. Güçlü, X. Baró, I. Guyon, J. C. Jacques, M. Madadi, *et al.*, "Modeling, recognizing, and explaining apparent personality from videos," *IEEE Transactions on Affective Computing*, 2020.

[52] G. Gunaydin and E. Selcuk, "The Role of Responsiveness Variability and Average Responsiveness in Relational Outcomes," 2018. The Science and Technology Institute of Turkey (TUBITAK) 1001 Scientific and Technological Research Projects Funding Program, Project no: 118K162.

[53] A. Dehghani, O. Sarbishei, T. Glatard, and E. Shihab, "A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors," *Sensors*, vol. 19, no. 22, p. 5026, 2019.

[54] C. S. Silveira, J. S. Cardoso, A. L. Lourenço, and C. Ahlström, "Importance of subject-dependent classification and imbalanced distributions in driver sleepiness detection in realistic conditions," *IET Intelligent Transport Systems*, vol. 13, no. 2, pp. 347–355, 2019.

[55] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 59–66, IEEE, 2018.

[56] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

[57] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

[58] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

[59] C. X. Ling, J. Huang, and H. Zhang, "AUC: A better measure than accuracy in comparing learning algorithms," in *Conference of the Canadian Society for Computational Studies of Intelligence*, pp. 329–341, Springer, 2003.

[60] K. H. Zou, A. J. O'Malley, and L. Mauri, "Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models," *Circulation*, vol. 115, no. 5, pp. 654–657, 2007.

[61] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[62] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long-and short-term temporal patterns with deep neural networks," in *International ACM SIGIR*

Conference on Research & Development in Information Retrieval, pp. 95–104, 2018.

[63] G. Gunaydin, E. Selcuk, and V. Zayas, "Impressions based on a portrait predict, 1-month later, impressions following a live interaction," *Social Psychological and Personality Science*, vol. 8, no. 1, pp. 36–44, 2017.