VIDEO BASED AUTOMATIC AFFECT ANALYSIS OF A CHILD AND A THERAPIST DURING PLAY THERAPY

by

Batıkan Türkmen

B.S., Computer Engineering, Bilkent University, 2015

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2019

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor, Assoc. Prof Albert Ali Salah. I count myself very lucky to be one of his students. Additionally, I want to express my sincere gratitude for his great advice, excellent guidance, and support during my thesis work.

I would also like to thank Prof. Dr. Lale Akarun, Assist. Prof. Sibel Halfon and Assoc. Prof. Arzucan Özgür for their valuable feedbacks and being a member of my thesis committee.

I would like to thank their motivation and support of my closest friends, Metehan Doyran, Burak Demirel, Ahmet Murat Artuç, Ahmet Yasin Tekin, and Gülbahar Eda Erbaş. I would also like to thank the members of Media Laboratory, Özlem Şimşek, Alper Ahmetoğlu, Çağla Aksoy, Oğulcan Özdemir and, Özlem Salehi Köken for providing a nice working environment.

Last but not least, I owe my deepest gratitude to my parents, Sevgi and Ümit, and my brother Kaan. This thesis would not be possible without their support, help, and patience.

ABSTRACT

VIDEO BASED AUTOMATIC AFFECT ANALYSIS OF A CHILD AND A THERAPIST DURING PLAY THERAPY

With recent developments in machine and deep learning techniques and the advent of big data, computer vision supports many disciplines, including social sciences. Although computer vision is used in social signal processing in psychology and its sub-branches, there is a lack of studies in the field of play therapy. Play therapy is a psychotherapeutic method, and it is a convenient yet challenging field to apply automatic computer analysis techniques due to the extensive range of body poses, the existence of various play activities, and occlusions with people and objects. In this thesis, we investigate an approach to track the affective state of a child during a play therapy session with two diagonal cameras. Moreover, to differentiate the therapist and the child in the therapy room, we introduce a human tracking module. Finally, we provide a web-based affect analysis tool for the field experts to interactively visualize affect over longitudinal data. We conduct experiments on various modalities and their fusions, including text analysis, face analysis and body motion analysis during the therapy session. In this study, we used about 350 hours of therapy videos, containing two million facial expressions. Our results show that the proposed system is promising and yet still open to improvement at different stages.

ÖZET

BİR ÇOCUK VE BİR TERAPİSTİN OYUN TERAPİSİ SIRASINDAKI DUYGU DURUMUNUN VİDEODAN OTOMATİK ANALİZİ

Makine ve derin öğrenme konusundaki gelişmeler ve büyük verideki ilerleyiş ile birlikte birlikte bilgisayarlı görme, sosyal bilimler de dahil olmak üzere birçok alanı desteklemeye başlamıştır. Her ne kadar psikoloji ve alt dallarında sosyal sinyal işleme için bilgisayarlı görme kullanılsa da, oyun terapisi alanında bilgisayarlı görme kullanımında çalışma eksikliği vardır. Psikoterapik yaklaşımın bir çeşidi olan oyun terapisi bu tür öğrenme tekniklerinin uygulanması açısından uygun fakat çeşitli vücut pozlarının varlığı, farklı oyun aktiviteleri olması ve insanlar ve objelerin üstüste gelmesi gibi nedenlerden dolayı zorlu bir alandır. Bu çalışmada, iki köşegen kamera kullanarak çocuğun oyun terapisi sırasındaki duygulanım analizini izlemek için bir yaklaşım araştırılmaktadır. Ayrıca, terapi odasındaki çocuğu ve terapisti ayırt edebilmek için geliştirdiğimiz insan takip modülü sunulmaktadır. Son olarak, alan uzmanları için zamansal veriler üzerindeki etkiyi etkileşimli olarak görselleştiren, web tabanlı duygulanım analizi aracı tanıtılmaktadır. Bu çalışmada, terapi seansı sırasındaki metinlerin, yüzlerin ve vücut hareketlerininin analizi olmak üzere çeşitli yöntemler ve bu yöntemlerin birleştirilmesi üzerinde deneyler yapılmaktadır. Bu çalışmada, iki milyon yüz ifadesi içeren yaklaşık 350 saatlik terapi videosu kullandık. Sonuçlarımız önerilen sistemin umut verici ve yine de farklı aşamalarda gelişime açık olduğunu göstermektedir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

B_k	Control matrix
F	Prediction matrix
H	Sensor matrix
K	Kalman gain matrix
P_k	Covariance of kalman estimates at time k
Q	External noise
R	Sensor noise
$\overrightarrow{u}$	Control vector
$\hat{x}_k$	Kalman estimates at time k
$\overrightarrow{z_k}$	Observation mean at time k
μ	Mean
π	Pi number
σ	Variance
$\sum$	Covariance matrix

LIST OF ACRONYMS/ABBREVIATIONS

2D	Two Dimensional
3D	Three Dimensional
API	Application Programming Interface
AU	Action Unit
CBCL	Child Behavior Checklist
CNN	Convolutional Neural Network
CPTI	Children's Play Therapy Instrument
CPU	Central Processing Unit
ELM	Extreme Learning Machine
EKF	Extended Kalman Filter
FACS	Facial Action Coding System
GMM	Gaussian Mixture Model
GPU	Graphics Processing Unit
HMI	Human Machine Interaction
HSV	Hue Saturation Value
JSON	JavaScript Object Notation
KF	Kalman Filter
MP4	MPEG-4 Part 14
MPEG	Moving Picture Experts Group
MSE	Mean Squared Error
NLP	Natural Language Processing
NMS	Non Maximum Suppression
PAF	Part Affinity Field
POS	Part Of Speech
RGB	Red-Green-Blue
SLFN	Single-hidden-layer Feed-forward Network
SVM	Support Vector Machine
SVR	Support Vector Regressor

VAD	Valence-Arousal-Dominance

1. INTRODUCTION

1.1. Motivation

Emotion is a significant element for communication, and visual cues during an interaction provide essential clues for the affective states [1]. Play therapy provides a natural environment in which children can express core emotions and social signals through play. In play therapy, the therapist draws the attention of the child to the play process by listening actively and inviting the child to communicate in play, encouraging the children to express themselves in the manner of his or her feelings, perceptions, and thoughts. The therapist supports the child reflect on the play context and the revealing emotions by asking questions about the play, and the details of the characters, their thoughts, feelings, and behaviors in terms of mental states. The therapist also helps the child understand the links between emotions about self and others that find reflection in play behaviors and the therapeutic relationship, to bring feelings, assumptions, attitudes, and beliefs into consciousness. All these interventions help the child organize and have a better understanding of their internal emotional world [2].

Although the significance of affect expression and mutual affective sharing processes in psychodynamic child therapies, the micro affective processes that take place within the sessions and associations with the outcome have not been thoroughly studied because, affective analysis of psychodynamic play therapy sessions is a meticulous process that needs many passes over the gathered data to annotate different perspectives of play behaviors and signs of affective displays. Furthermore, there is a need to combine multiple modalities that take into account verbal and nonverbal indicators of affect for a comprehensive affect assessment. Moreover, infrastructure is demanded to examine not only treatment but also longitudinal affect outcomes.

Automatic assessment of affective states in play therapy videos is a challenging task because of the extensive range of body poses, the occurrence of various play activities, and occlusions with objects and people. However, most importantly, chil-

dren's expressions of emotions are highly idiosyncratic and show a notable variation depending on the context.

In this thesis, we investigate an approach to track the affective state of a child during a play therapy session. We evaluate off-the-shelf deep convolutional neural networks for the processing of the child's face during sessions to automatically extract valence and arousal dimensions of affect, as well as basic emotional expressions. Moreover, we examine the text and body-movement based affect analysis, and evaluate these modalities separately and in conjunction with play therapy videos in natural sessions, discussing the results of such analysis and how it aligns with the professional clinicians' assessments. Further, we investigate children with internalizing or/and externalizing problems and show disease prediction accuracy from the modalities. In the end, we will introduce the affect analysis tool, which allows therapy to be monitored on a session-based, play-based, or as a time series.

1.2. Related Work

Recent progress in affective computing proves that machine and deep learning strategies for affect analysis could be used to produce models that provide similar outcomes with expert assessments [3–5]. Trained models can achieve high reliability and infer emotional states relatively better than humans in controlled recording environments [6]. On the other hand, there is also evidence that states that contextual signals actively shape the interpretation of emotional expressions [7]. Expert humans can consider these better than computer systems and are much better in distinguishing rare events and subtle changes. However, training human experts is expensive, and each new annotation on the data requires time and effort, it is favorable to understand the boundaries and capabilities of automatic analysis systems.

We work on play therapy that is a well-known methodology applied for children with behavioral and emotional to help them to be able to control their problematic behaviors. Play is a wealthy context for investigating the affective state of the children [8, 9]. Halfon *et al.* investigate the relation between affect and play therapy as

well as the assessment of affect in the therapy [10] and proposed a natural language processing based solution for automatic affect analysis during play. Nevertheless, facial and bodily expressions are the most significantly used signals for automatic affect analysis [11, 12]. Poria *et al.* showed a detailed overview and presented the benefits of analyzing videos for capturing the affective states, endorsing the reliability of facial expressions and visual data cues.

Affect analysis of play can serve human knowledge in different ways. Robot-assisted autism therapy is an approach that hired humanoid robots, tries to create an individual robot-child interaction during the therapy, having the essential purpose of capturing facial images from cameras placed onto the robot, and assessing the emotion of the captured faces. Using robots is interesting for a child and could be beneficial for therapy [13, 14]. An interactive photo frame is another usage of tracking the affect of the child with the facial features. For interactivity authors assess the affect on a person's face, who looks to a frame that is showing videos and presents about the similarity of affect scores [15]. Affect analysis use not only visual features, but also non-visual features can imply used. A simple rule-based natural language processing system uses the transcripts of the play can also extract valence, arousal and dominance scores of the therapy sessions, and can give clinicians a chance to detect internalizing and externalizing problems of a child using the transcripts of the play [10].

Facial expression analysis for children is a beneficial tool for settings beyond play. Autism can be detected with expression analysis. The works show that autistic kids have fewer complex dynamics around their eye regions [16]. On the other hand, analyzing signals to detect deception in children [17] is a challenging task. The insights of children's social behavior can be revealed by processing facial expression, head pose, and gaze [18]. There is a need for new databases centering on children's expressions. Khan *et al.* introduced LIRIS-CSE for children's spontaneous expression recognition [19]; however, it contains videos from 12 children, and it is not enough to train classifiers that can generalize adequately. There is also a need for tools that can deliver convenient feedback to the domain experts.

In this thesis, we do not focus analysis of the speech emotion. However, there are methods that specialize in recognizing emotions in children's speech as well. The FAU Aibo Corpus consisted of 8.9 hours of audio recordings of emotional children's speech in German during interactions with a robot, including Emphatic and Anger classes [20], which used for an Interspeech Challenge in 2009, accelerated the area research in this course [21]. Also, Lyakso *et al.* introduced the EmoChildRu corpus, for emotional speech that collected from 100 children between three and seven years old [22]. Emotion detection across languages does not work well, and the training of such systems should be with the same language as the testing unless elaborate normalization techniques are used [23]. Speech emotion detection resources are very limited for Turkish language [24].

Children have different emotion characteristics compared to adults and their emotions differently both in terms of speed and variability [25]. According to Monier *et al.* speed and variability can change depending on the cooperation between child and adult. In our dataset, children with different emotional and behavioral problems (i.e., internalizing, externalizing, and co-morbid) show different levels of affect creation and cooperation. Our experimental setting also includes children with no known emotional or behavioral disorders, which results in a highly variable setup in terms of affect production.

Videos of interactions that are subsequently rated by experienced coders have been a viable alternative to self-report based measurements, which are difficult to use with children [26, 27]. Since expert coding of facial expressions is an expensive task [28], computational alternatives were developed to recognize action units from videos [29, 30]. Recently, progress in deep learning, combined with access to a vast amount of the face datasets caused significant improvements in the robustness and accuracy of these approaches [31, 32].

Children with special disorders require additional expertise in creating automated tools of analysis for helping therapists during their therapy procedures [33] or diagnosis processes [34]. Some degree of tailoring to the special needs of the group seems essential.

In our system, we were adapting domain-specific words and their affective value.

1.3. Contributions

We developed an affect analysis pipeline and three main contributions can be listed during the development of the pipeline. Our first contribution is suggesting a multimodal affect analysis system. The second contribution is the skeleton tracking module which is to track people in the therapy room, and our last contribution is designing and implementing affect analysis tool to monitor and store affective inferences from the children's therapy.

We present a novel multimodal system to evaluate affective expressions of children and their therapists' emotional responses in the play therapy. Automated evaluation methods are especially significant for continuous monitoring because survey-like assessments cannot be applied to the patient frequently. If we consider the amount of information produced by each child undergoing regular therapy, the annotation effort is significantly costly. We aim to facilitate the work of the therapists in their evaluation. Moreover, we want to support retrieval-style scenarios by producing appropriate indices and minimizing the faults that may occur from an annotator by using an automated system.

Although capturing a clear face in the video is essential for facial analysis, it is not possible to ensure this during the play of children, even if multiple cameras are used simultaneously. An extreme scenario is the work of Joo *et al.* [35], where 480 synchronized cameras were used to record the interactions simultaneously, in what is called a Panoptic Studio. However, such a system would be extremely expensive to setup and maintain, and produce a great amount of data to process. Using multiple cameras increases the setup and running costs of the systems, and typically one or two cameras are used in recording scenarios, and rarely more than six.

Since we work with legacy data, we also face this challenge, and try to overcome it through multimodality. The dataset we work with was collected with two cameras,

and clear face images are difficult to obtain. We subsequently propose a multimodal method that combines facial affect analysis with language-based affect analysis, which overcomes this limitation up to a point. Also, we developed an automatic tracking module so that we can track the child and the therapist during play therapy, to be able to perform facial analysis.

Lastly, automatic facial and lexical affect analysis tools were developed, adapted, and combined in a system that interactively visualizes affect over longitudinal data. The tools can be used for representing and indexing a large amount of aggregated session data, to reveal patterns and trends, to visualize affect dynamics, and to measure correlations between the automatically extracted expression streams of the therapist and the patient through sessions.

We publish our findings of affect analysis from a child's face during therapy as a poster to the International Symposium on Brain and Cognitive Science (ISBCS) 2019. Furthermore, we submit two journals which are Journal of Counseling Psychology (JCCP) 2019, where we made clinical study through automatic affective analysis of children with the face and text based affect analysis and their fusions, and International Conference on Multimodal Interaction (ICMI) 2019, where we made CPTI predictions by using face and text based affect analysis and play area estimation.

1.4. Organization of the Thesis

The rest of this thesis is organized as follows. In Chapter 2, we provide some domain knowledge on play therapy and give a brief introduction about emotion, affect and motion measurement methods, and mental disorders. We also summarize state-of-the-art facial expression data sets, and we explain the data we use in this study in detail in Chapter 2. In Chapters 3, 4, and 5, we describe the evaluation of our affect analysis pipeline and show building blocks and development process of the approaches to achieve the proposed methodology, which are used to obtain our results in depth. We also report and discuss our findings for each approach in these sections. Furthermore, we introduce our affect analysis tool and its capabilities in Chapter 5 Finally, we

conclude our thesis work in Chapter 6 and provide some discussion about the larger implications of using automatic analysis tools in psychology and psychotherapy.

2. BACKGROUND AND DATA

In this chapter we give brief introduction about psychological background of the work. After making brief introduction about the emotion and the valence, we mention about the affect analysis databases. Finally, we will explain the dataset and metrics that we use to develop and measure our affect analysis methodology. It is important to emphasize that any visual or identity information about children will not be shared in the thesis in terms of protecting the identity of children and the privacy of the doctor-patient confidentiality.

2.1. Emotion and Affect

Emotion has been investigated under several disciplines through both philosophical and scientific approaches. Oxford dictionary explains emotion as 'instinctive or intuitive feeling as distinguished from reasoning or knowledge.' The difference between emotion and affect should be well explained to clear the question marks. Affect is the psychological term that is different from emotion and explains the experience of the emotion. Emotion can be modeled with the categorical model, dimensional model, and action units (AU).

In categorical modeling, emotion has classes, and most significant work on emotion categorization is introduced by Ekman *et al.* [37] which defines six main emotion categories shown in Figure 2.1 which are happiness, surprise, sadness, fear, disgust, and anger. Affect categorization generally grounds on Ekman's six basic emotions. However, there are alternative approaches to extend these categories. For instance, Plutchik *et al.* claims that basic emotions are anticipation, acceptance, anger, disgust, sadness, joy, fear, and surprise [38] and Izard *et al.* state that there are ten basic emotions which are disgust, anger, contempt, joy, fear, guilt, interest, surprise, distress, and shame, respectively [39]. The categorical model is limited due to a certain amount of emotion classes. Some works use categories together to cope with the limitation problem and increase emotion classes to 21 [40]. For instance, happily surprised is the

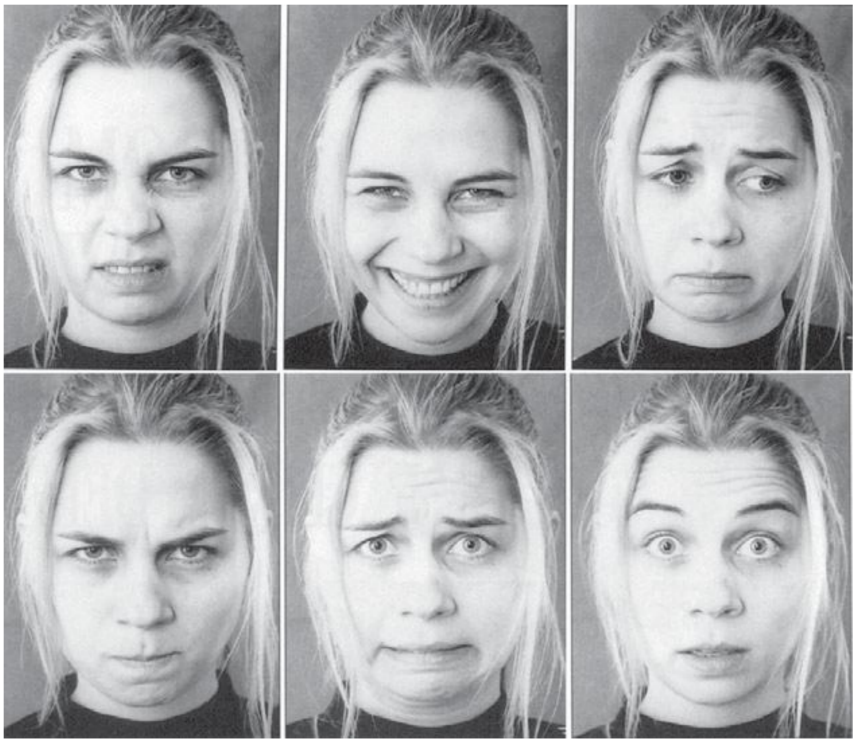

Figure 2.1. Ekman's six basic emotions. In order (left-to-right, top-to-bottom) Disgust, Happiness, Sadness, Anger, Fear and Surprise [36].

combination of being happy and being surprised, and can be shown in Figure 2.2. However, this approach does not provide the desired flexibility due to the limited number of classes.

[41] In dimensional modeling, the affective state can be shown in a multidimensional space. The first dimensional approach is introduced by Wundt *et al.*, who used three dimensions to describe emotion [42]. These dimensions represent pleasure, arousal, and relaxation. Russell *et al.* proposed a circumplex model for affect representation and places the emotion categories on the circumplex [43]. In the circumplex model, emotion categories are placed in a 2D multidimensional space, where the axes represent valence and arousal. Valence measures how negative or positive the feeling of emotion is, and arousal measures the energy in the experience. The origin corresponds to the neutral emotion in the model. Verma *et al.* extending this model by adding dominance as a third dimension, which represents being submissive or dominant [41]. Valence and arousal have continuous values, so that even slight changes in the affective

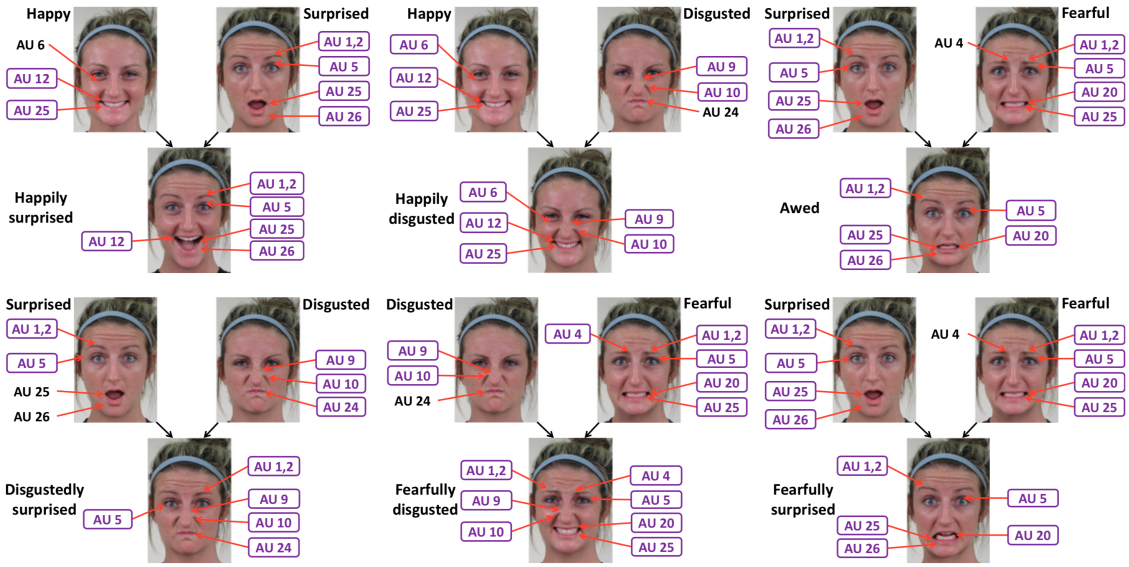

Figure 2.2. Compound facial expressions of emotion [40].

state can be encoded thanks to this continuity. On the other hand, measuring and coding this slight change prune to the error for both human experts and the computer aided systems. Dimensional values can be represented in the coordinate system. An example regarding the representation of valence and arousal is given in Figure 2.3.

There is also another approach to measure facial characteristics which is facial action coding system (FACS). In the facial action coding system (FACS), facial muscle movements are coded as Action Units [44]. The action unit does not directly show the emotion; however, there is a study that shows that affective states can be coded with the combination of AUs [45]. In the action unit coding system, for example, happiness is coded with AU6 and AU12, sadness is coded with AU1, AU4, and AU15, and fear is coded with AU1, AU2, AU4, AU5, AU7, AU20, and AU26. However, action unit representation cannot express how intense the emotion is.

Human affects can be recognized by computers, which creates a subfield for computer science called affective computing, which is introduced in 1995 by Rosalind Picard [46]. Although most of the affective computing studies engage in determining the affect from facial expressions, there are studies which try to make inferences from speech [47], postures [48], and physiological signals [49]. In the early stages of affective

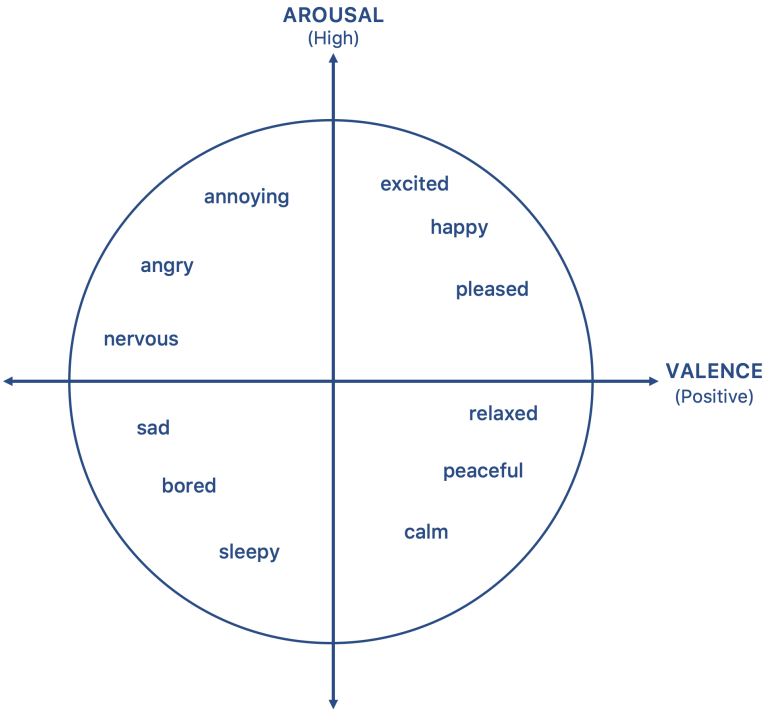

Figure 2.3. Representation of the dimensional model of affect.

computing, approaches have focused on recognizing the emotion categories with the data which collected highly controlled conditions in the manner of illumination and head pose. However, current studies focus on recognizing dimensional affect [50] under in-the-wild conditions [51].

2.2. Measurements

2.2.1. The Children's Play Therapy Instrument

Children's Play Therapy Instrument (CPTI) [52] is used to evaluate the dynamics of the play and to determine a criterion to measure the proposed approach. During the therapy, only play sections of the session is scored. CPTI contains various measurements that can be shown in Figure 2.4. On the other hand, in this work, we use only affect and sphere classes. The affect classes show how much the child reveals the emotions during the play, and the sphere classes specify the area of play in the therapy room. Both affect and sphere class indices are scored between zero and five.

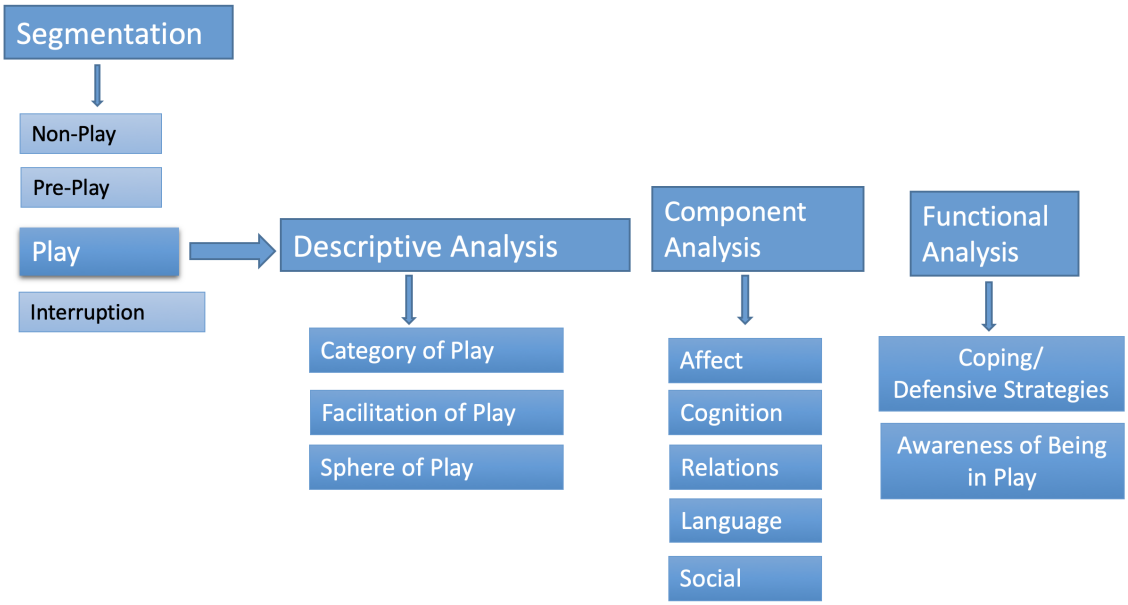

Figure 2.4. Children's play therapy instrument (CPTI) categories.

We use two different versions of CPTI in this work. The first version of CPTI has eight affect classes, which are anger, anxiety, fear, boredom, pleasure, sadness, shame, and guilt. The second version of CPTI has four affect classes, which are anger, anxiety, pleasure, and sadness. Moreover, the second version of CPTI contains the score with 0.5 step size and also none values which indicate that the corresponding affect has not been shown during the play, and we consider all none values as zero.

Both CPTI has two sphere classes that are microsphere and macrosphere that use the same method of measurement. Sphere classes indicate how the child uses the space of the room during the play. Macrosphere is the measurement to evaluate how the child dominates the room during the play. For instance, if the child play ball games or dart, s/he dominates the entire room. On the other hand, microsphere is an assessment for play in small areas such as small toys or painting. The significant point is that for a single play, both scores can be high or low together. However, these values are highly correlated in our dataset with -0.82 correlation.

2.2.2. Child Behavior Checklist

The Child Behavior Checklist (CBCL) [53] helps the evaluation of problematic behaviors in children. For age 1.5-5 and 6-18, two different versions of CBCL is applied. CBCL uses three point scale as not true, somewhat true, and often true, and CBCL consists of 112 questions that need to be answered with this point scale. Through analyzing CBCL outcome, internalizing (e.g., depression, anxiety), externalizing (e.g., aggression, violence) or total problems can be diagnosed. The native language of CBCL is English, yet it has been adapted to Turkish [54].

2.3. Internalizing and Externalizing Problems

Mental problems in childhood can be examined under two main categories, which are internalizing and externalizing. Children with internalizing disorders show over controlled behavior such as anxiety, fear, and depression. On the other hand, children with externalizing problems show under controlled behavior such as attention deficit hyperactivity disorder, aggression, and conduct disorders [55]. Furthermore, the child can be comorbid, which means to show both internalizing and externalizing problems. Comorbid children have difficulties in regulating themselves against negative emotions [56]. In other words, anger is generally related to externalizing symptoms, whereas dysphoric affect such as fearfulness, anxiety, and depression are results of internalizing problems [57].

Children with externalizing or internalizing problems have different play characteristics. A child with externalizing problems shows more negative affects, such as aggression in the play [10]. On the other hand, a child with which has internalizing problems shows high negative affects and low arousal [58]. These children also play solitarily and have less organization in play [59].

2.4. Affect Databases

As in all subjects of machine learning and deep learning, data is very crucial in the learning process. Collecting data sets for emotion analysis proceeds with exponential growth for many years. These datasets range from black and white face data through color video data to electrocardiography [60–62] with different subject sizes. A brief overview can be seen in Table 2.1.

In this thesis, we used the state-of-the-art database with most data points called AffectNet [76]. In our works, we want to take advantage of not only the emotion categories but also intensity to be able to track affective states of the people in the therapy so that AffectNet meets our requirements. Table 2.2 shows the distribution of the emotion classes of AffectNet, and sample images with their valence and arousal distribution on circumplex can be seen in Figure 2.5.

In the creation of AffectNet, three different search engines were queried with 1250 different queries in six different languages, which are English, Spanish, Portuguese, German, Arabic, and Farsi. AffectNet contains 11 categories, which are neutral, happy, sad, surprise, fear, anger, disgust, contempt, none, uncertain, and non-face. If an image does not contain a face or contains watermarks on it, it is assigned to the non-face category. If there is a contradiction about what emotion belongs to a face, it is labeled as uncertain. If an image has an expression other than these eight categories such as confuse, shame, sleepy, tired, bored, and so on, it is considered as none of the eight emotions with none label. On the other hand, valence and arousal could be assigned to these images. Images with non-face or uncertain labels do not contain valence and arousal.

Table 2.1. Overview of affect datasets.

Database	Year	Information	# of Subjects	Shape	Affect Modeling
Cohn-Kanade (CK+) [63]	2010	- Frontal and 30 degree images	123	- Controlled - Posed	- 30 AUs - 7 emotion categories
MultiPie [64]	2010	- ∼750,000 images	337	- Controlled - Posed	- 7 emotion categories
MMI [65]	2005	- Frontal and side face	25	- Controlled - Posed & Spontaneous	- 31 AUs - Six basic expression
DISFA [66]	2013	- Recorded in lab environment while watching video - Stereo camera	27	- Controlled - Spontaneous	- 12 AUs
SALDB [67]	2010	- Audiovisual	4	- Controlled - Spontaneous	- Valence
RAFD [68]	2010	1005 Posed face images	67	- Controlled	- 8 emotion categories
RELOCA [62]	2013	- Multi-modal audio, video, electrocardiography	46	- Controlled - Spontaneous	- Valence and arousal
AM-FED [69]	2013	- 242 videos	242	- Spontaneous	- 14 AUs
DEAP [70]	2011	- Frontal face videos- EEG signals recorded	32	- Controlled - Spontaneous	- Valence and arousal - Self assessment
AFEW-VA [71]	2017	- From movies	600	- Wild	- 7 emotion categories - Valence and arousal
FER-2013 [60]	2013	- Greyscale images - Web	35,887	- Wild	- 7 emotion categories
FER+ [72]	2016	- Having same data with FER-2013 - Relabeled with 10 taggers	35,887	- Wild	- 7 emotion categories
Aff-Wild [73]	2017	- 298 videos - YouTube	200	- Wild	- Valence and arousal
GIFGIF+ [74]	2017	- 23,544 emotional GIFs - Web	23,544	- Wild	- 17 emotion categories
FER-Wild [75]	2016	- 24,000 images - Web	∼24,000	- Wild	- 7 emotion categories
LIRIS-CSE [61]	2019	- Videos - Children's spontaneous facial expressions	12	- Wild	- Six basic expression
AffectNet [76]	2017	- 1,000,000 images - Web	∼450,000	- Wild	- 8 emotion categories - Valence and arousal

Table 2.2. Distribution of labeled emotional expressions in the AffectNet.

Emotion	# of Child Faces
Neutral	80,276
Happy	146,198
Sad	29,487
Surprise	16,288
Fear	8,191
Disgust	5,264
Anger	28,130
Contempt	5,135
None	35,322
Uncertain	13,163
Non-Face	88,895

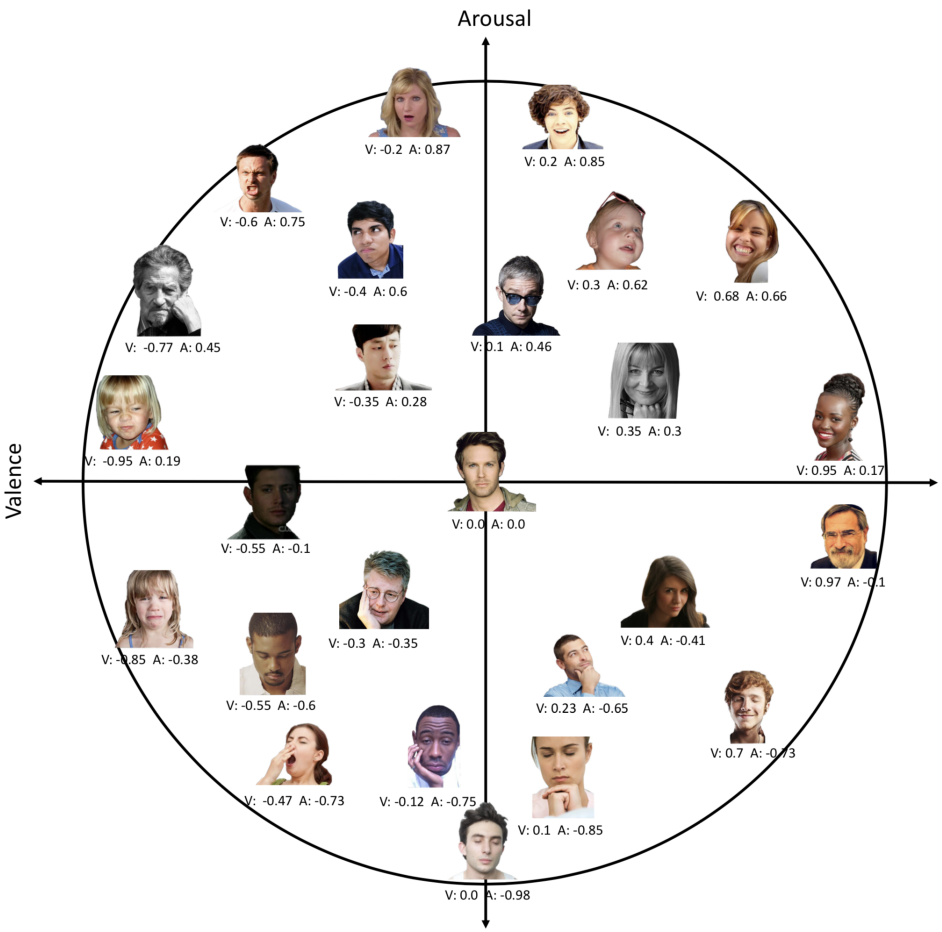

Figure 2.5. Sample images in valence arousal circumplex [76].

2.5. Play Therapy Dataset

In this work, we use the play therapy video records which are provided by Istanbul Bilgi University in order to research the psychotherapy sessions of children. Provided videos are recorded simultaneously by two cameras located diagonally at the corners of the therapy rooms, and one of these cameras also provides audio information. The videos were recorded with 25 frames per second and 1280×720 resolution in MP4 format. Moreover, to be able to get better sound quality, a separate sound recorder also records the sounds in the room during the therapy. Conversations in all therapy sessions have been transcribed. Sample room photos can be seen in Figure 2.6.

Figure 2.6. Sample therapy room images from two different cameras.

Play therapy session videos contain four subcategories that are 'non-play', 'pre-play', 'play' and 'interruption'. Segments of session videos are annotated by profes-

sional clinicians, according to CPTI. Each session contains one or more play subcategories with varying lengths. Segments meanings are as follows: Non-play segment means that the child has done nothing about the play. In the non-play segment, the child either talks to the therapist or does not do anything. The pre-play segment means that the child intends to start playing, and the play segment means that the child plays. In this work, we only focus on the play segment. The interruption segment means that the child leaves the game and does another job, such as going to the toilet or drinking water. A total of 645 play segments from 310 sessions of 52 children are used in this work.

2.5.1. Patient Characteristics

The videos are recorded in Istanbul Bilgi University Psychotherapy Research Laboratory, and the aim of the laboratory is to provide low-cost psychodynamic psychotherapy.

The children in this thesis were randomly selected based on data integrity from 90 children who admitted from Fall 2014 to 2017 and met criteria: 4-10 years old, has no psychotic symptoms, no significant developmental delays, no significant risk of suicide attempts and no drug abuse. The integrity of the data is ensured by the fact that the transcripts are complete, both two cameras contain video, and the CPTI values are not missing. Before the treatment, patients and their parents were informed about the research procedures, and the parents gave written informed consent to the use of their data, and this research was approved by Istanbul Bilgi University Ethics Committee.

The children in the dataset were born in Turkey, and Turkish is their mother tongue. The children belonged to low to middle socioeconomic status and came from urban neighborhoods. 26% of the children were 4–5 years old, 28% were 6–7 years old, 46% were 8–10 years old). 69% of the sample was female. They were referred most frequently due to internalizing and externalizing problems such as rule-breaking and aggressive acts (48%), followed by anxiety complaints (26%), school-related problems (19%) and social problems (7%). 11% of the children had internalizing problems, 11%

had externalizing problems, and 57% had comorbid internalizing and externalizing problems according to CBCL.

2.5.2. Therapists

44 graduate-level clinicians are in the dataset, and these therapists are between the ages of 23 and 27 and mostly female (95%). Each therapist was trained in the theoretical background of psychodynamic game therapy, which was informed by the principles of mentalization for two years in theoretical courses [77]. All therapists have supervised psychotherapy experience about one to two years, and one therapist accounts for the treatment of about three patients. Therapists are educated by supervisors with at least ten years of experience.

2.5.3. Treatment

Psychodynamic play therapy is applied in the Istanbul Bilgi University psychotherapy center. The therapy follows an object-relational framework that uses children's play as the main source of internal expression, focusing on the play of children with their inner representations and the associated mental states [77]. Children are assigned to therapists according to the availability of them. Standard therapy is done once a week and lasts 50 minutes. There is also a session with the family of the child once a month. Treatments are shaped based on changes, goals, family decisions of the patients. On average, the child receives 40 session therapies in more than ten months period.

2.5.4. Dataset separation

The CPTI assessment methodology has changed after 2015. Previous assessment methodology scored each play in every session for each child. After the change, only the longest play in the session is scored. Furthermore, only one session was randomly chosen from sessions 1-10, 11-20, 21-30, 31-40, 41-50 for each child. These changes cause unbalance in the data; thus, we have to use two different combinations of the

data. To avoid confusion, we assign names to these two different datasets. Due to that shape of the data, we called the dataset vertical, which has fewer children, but more play and the one with more children and fewer plays horizontal. The detailed comparison can be shown in Table 2.3. In addition, we have ten children marked with old methodology and 42 children after the methodology change. This change causes a loss of data. On the other hand, since we own all data of 10 children, we perform data selection to minimize our data loss and convert it to the same data structure as the other 42 children, so that after the change we have 52 children without losing any.

Table 2.3. Comparison of play datasets

Dataset	Vertical	Horizontal	Intersection
# of Children	10	52	10
# of Session	183	151	24
# of Labeled Play Segments	391	302	48
CPTI Emotion Count	8	4	8
Annotated Plays	All plays in the session	Longest play in the session	Longest play in the session

3. COARSE AFFECT ANALYSIS AND RESULTS

In the first methodology, our aim is to find a meaningful association between the CPTI and facial affect of the children to understand whether there is a relationship between CPTI and facial affect. In this work, we use Vertical Database that is explained more detailed in Section 2.5. Sample snapshots from the play therapy environment can be seen in Figure 3.1.

Figure 3.1. Sample scenes from the play therapy dataset, processed with a style transfer neural network to preserve the privacy of the participants.

We use OpenPose to detect faces in play therapy videos [78]. OpenPose is a tool to detect human body and the facial landmarks and information about the working principle of OpenPose can be found in 3.2.The facial images that are extracted from OpenPose are fed to a deep neural network for recognizing valence and arousal dimensions of affect, as well as basic emotions [76]. Neural networks is a method in machine learning and its application areas can be ranged from classification to regression. More detailed explanation can be found in the Section 3.1. Figure 3.2 illustrates the overview of the system.

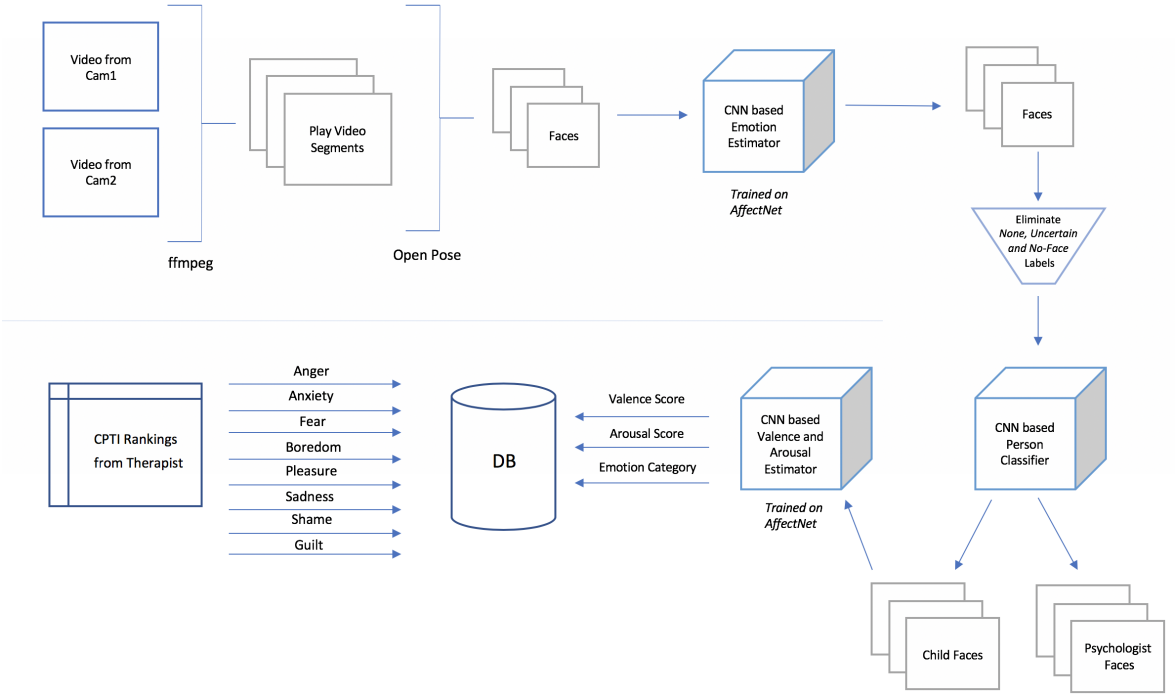

Figure 3.2. The schematic layout of the coarse affect analysis framework.

3.1. Neural Networks

The idea of neural networks is the result of an attempt to model the human brain. In the computational model, neurons take input and produce output like neurons in the human brain. Each neuron multiply received input with its weights, add the bias, and use activation function F to determine whether the neuron is fired. Sigmoid function 3.1 is one of the commonly used activation function, and it takes the input and limits it between zero and one with respect to their rate.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.1}$$

The connection of neurons forms neural networks, and the neural networks consist of any number of layers and any number of neurons located on these layers. Figure 3.3 shows a 3-layer neural network with three inputs. To perform training on the

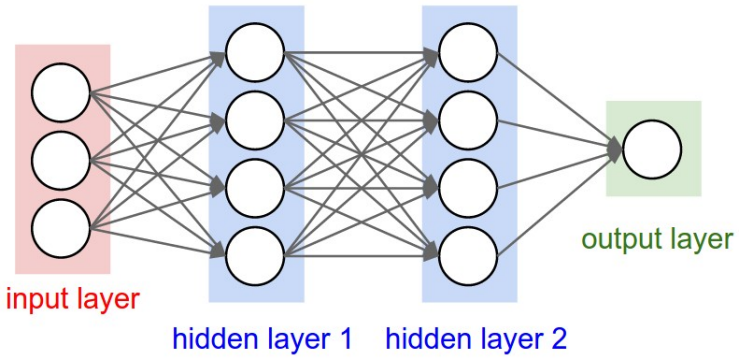

Figure 3.3. A three layer neural network with three inputs [79].

neural network, we should say how the prediction is far from the real result, and loss function is used for this metric. Neural networks can use various loss functions and mean squared error (3.2) is one of them. In other words, training the neural network means minimizing its loss.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_{true} - y_{predicted})^2 \tag{3.2}$$

To be able to update weights to minimize loss, neural networks used the technique called backpropagation (3.3). The main idea of the backpropagation is writing the loss as a multivariable function that consists of weights and biases and calculating partial derivatives to update variables.

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial y_{pred}} * \frac{\partial y_{pred}}{\partial h_1} * \frac{\partial h_1}{\partial w_1} \tag{3.3}$$

Neural networks have different types such as recurrent neural network (RNN), long-short term memory (LSTM), convolutional neural network (CNN). Each of these networks is derived from neural networks but using different approaches. Normal neural

networks assume image as a vector and treat each pixel as a single feature, but fully-connected layers cannot scale for larger images. On the other hand, CNN is better for image related tasks because it uses adjacent pixel information to downsample the image by convolution. Thanks to the downsampling fully-connected layers do not deal with an extensive amount of features.

CNN architectures use various layers, which are convolution, padding, max pooling, and flattening. In the convolution layer, the input image convolves with the filter to signalize the image. In the padding stage, padding is added to the image that shrinks the size due to the convolution operation. The maximum value is retrieved in the max pooling stage, and the flattening stage vectorizes the image. The Figure 3.4 shows example VGG 16 network architecture with its layers [80].

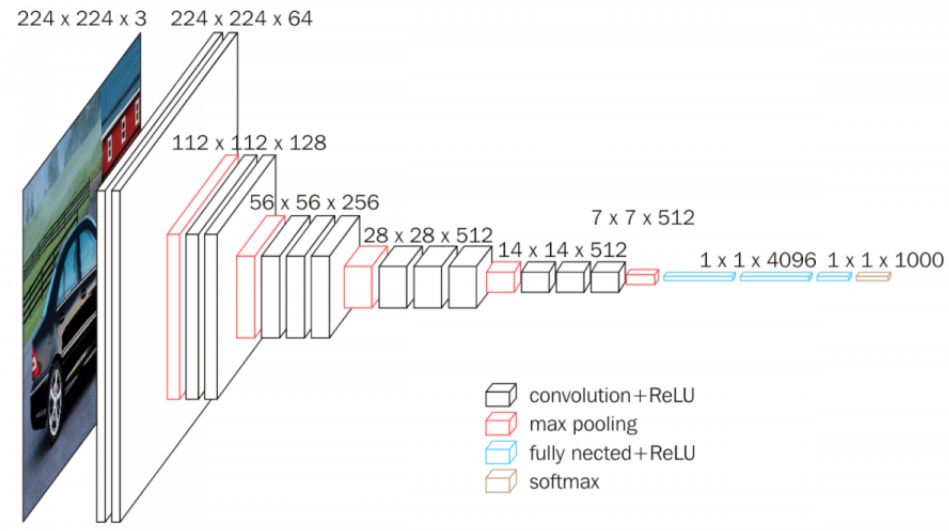

Figure 3.4. VGG 16 network architecture [81].

Transfer learning is a technique that transfers the learned knowledge from different problem to another problem. For instance, if a network trained to classify objects, we can change it to classify the emotions. Due to the computation limitations, and it is hard to find the data that has sufficient size, training an entire convolutional neural network from scratch is very costly. Therefore, it is common to use a pre-trained model and apply transfer learning on it.

There are several strategies to apply transfer learning. One technique is to remove the fully connected layer of the pre-trained neural network and treat the remaining network as a feature extractor. The linear classifier can be trained to classify extracted features. The second way is that not only training the classifier layers but also fine-tune the previous layer weights'. However, training upper layers of the network prevents over-fitting and reduces the processing load.

3.2. OpenPose

OpenPose is a tool that gives human skeleton data with parts and facial landmarks [78]. The output of OpenPose consists of 18 body landmarks and 70 facial landmarks, which can be seen in Figure 3.5. OpenPose uses the parts and pairs to create a skeleton on the given image. The part represents the body section such as hip and neck, and pair corresponds connection between two parts. Parts and Pairs can be seen in Figure 3.6.

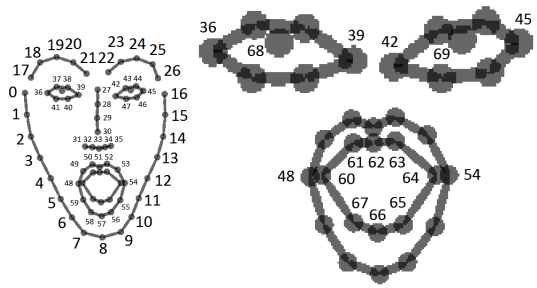

Figure 3.5. OpenPose facial landmarks [78]

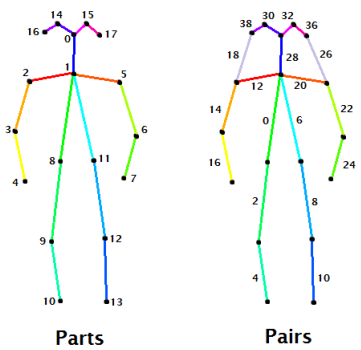

Figure 3.6. Index of parts and pairs in OpenPose [82]

The entire pipeline of OpenPose demonstrated in Figure 3.7, and it uses two separate deep neural networks to generate heatmap and part affinity field (PAF) [83]. One heatmap matrix is generated for each part of the body. The generated matrix indicates whether the corresponding region of the body is in that pixel with their confidence. In addition, for each pair PAF matrix is extracted for both 'x' and 'y' direction and PAF matrices store information about the position and the orientation of pairs that connect parts. Therefore, 18 heatmap and 38 PAF matrices are generated for a single frame.

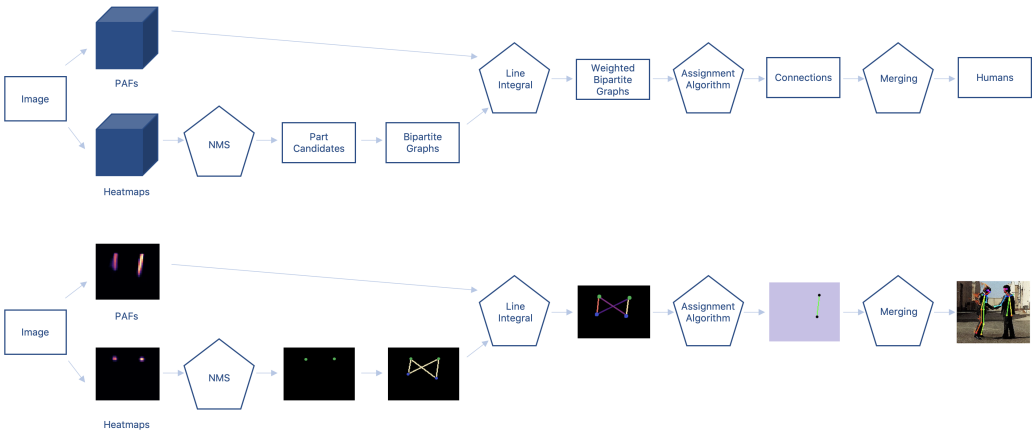

Figure 3.7. OpenPose pipeline.

After extracting heatmap matrices, confidence should be transformed to certainty. Non maximum suppression (NMS) is applied to heatmap matrices for transformation. NMS looks each pixel in matrix sequentially with a 5x5 window sized filter. Filter aims to find maximum in the window and subtract it from the center pixel to find peaks in the entire image. In the end, final results compared with original heatmap matrices and if pixel values are not changed, these are pixels that we want and suppress all other pixels by setting their values 0.

We need to find pairs to associate founded body parts with each other. We treat the whole body as a complete bipartite graph and try to find possible links. Moreover, graph edges represent connection candidates whereas vertices represent part candidates. This approach is similar to the assignment problem, and using PAF matrices, we found the weight of the graph. By taking the line integral of each part candidate,

we can find the weight of the graph edges. Taking line integral measures effect on PAF along feasible connections between candidates and the formula can be seen in Equation 3.4.

$$\int_{y_1}^{y_2} \int_{x_1}^{x_2} \begin{bmatrix} PAF_x(x,y) \\ PAF_y(x,y) \end{bmatrix} \cdot \begin{bmatrix} v_x \\ v_y \end{bmatrix} dxdy \tag{3.4}$$

The graph has weights for each possible candidate, and we should maximize the total score by solving an assignment problem for finding connections. To be able to solve the assignment problem, scores in the graph sorted from highest to lowest, and the highest score for each pair is selected. OpenPose assumes every pair is different human, and merging these pairs is a basic assumption that if there is a common part, the pair belongs to the same human. In the end, the skeleton for a human is exported with their confidence and location for the given frame.

3.3. Facial Feature Extraction

The deep neural network we have used for affect prediction produces an 11-class output for basic emotional expression estimation. These are neutral, happiness, sadness, surprise, fear, disgust, anger, contempt, none, uncertain, and no-face, respectively. Additionally, a second deep neural network is used, which produces real-valued valence and arousal scores. The used pre-trained deep neural network takes a square crop as input to process the facial images. Consequently, the cropped face area is square shaped. This network also removes images that are not sufficiently face-like (labeled as no-face and uncertain).

Faces are cropped with a margin with respect to the bounding box of the facial landmarks. Once the faces are located, their rotations are taken into consideration. OpenPose detection provides a set of landmarks for the face. We use ten stable land-

marks to determine the pose, and rotate the face to bring it to a frontal pose.

To be able to select the optimal crop size, we create a subset with ∼10000 facial images and try various approaches on it such as thresholding with respect to OpenPose facial landmarks accuracy mean different crop sizes and rotation. We used AffectNet pre-trained deep neural network for recognizing facial affect. To select optimum setup, we consider both AffectNet accuracy and output image count that do not includes None, Uncertain, and Non-Face. We also check outputs not only by looking at the scores but also by manually. In the end, we achieve the best results with 40% crop offset and without using OpenPose accuracy. All comparison results can be seen in Table 3.1.

We use a third deep neural network to separate child, and psychotherapist faces automatically [80]. We use transfer learning methods to fine-tune this network to separate the located faces into person classes.

Since we have two video recording for each play segment, and we would like to summarize the entire session in a single affect score, we use summarizing functionals to compare our features with CPTI. For each play segment, the obtained valence and arousal values are passed to the following functionals: mean, min, max, median, mode, standard deviation, variance, harmonic mean, range, mean absolute deviation, mean of top 10 highest value, mean of top 50 highest value, mean of top 100 highest value, mean of top 10%, mean of top 25%, mean of 10 smallest values, mean of 50 smallest values, mean of 100 smallest values, mean of smallest 10%, mean of smallest 25%, respectively. These 20 features were extracted for both valence and arousal, totaling 40 features per play segment.

For basic emotional expressions, a similar approach is followed. Each video is represented by a normalized distribution over eleven labels predicted by AffectNet, treated as a binned representation. We use both the number of frames selected for each label (resulting in 11 features) and the proportion of that label in the video (resulting in an additional 11 features). Less informative dimensions (such as no-face

and uncertain) are eliminated in the subsequent correlation analysis automatically.

We have investigated the correlations between the summarizing features and the CPTI affect labels assessed by the experts for each segment. There are eight affect classes in CPTI; aggressiveness/anger, anxiety/worry/wariness, fear, boredom, pleasure, sadness, shame, guilt, respectively.

Table 3.1. Comparison of face crop methods.

OpenPose Accuracy Threshold	Crop Offset	Rotation	AffectNet Accuracy	Output Image #	OpenPose Accuracy Threshold	Crop Offset	Rotation	AffectNet Accuracy	Output Image #
0.10	0%	No	0.419	960	0.35	0%	No	0.420	374
0.10	0%	Yes	0.426	583	0.35	0%	Yes	0.431	207
0.10	10%	No	0.382	2036	0.35	10%	No	0.365	541
0.10	10%	Yes	0.387	1523	0.35	10%	Yes	0.369	319
0.10	20%	No	0.361	3965	0.35	20%	No	0.358	1090
0.10	20%	Yes	0.358	3580	0.35	20%	Yes	0.361	911
0.10	30%	No	0.336	4132	0.35	30%	No	0.347	1547
0.10	30%	Yes	0.341	4001	0.35	30%	Yes	0.351	1455
0.10	40%	No	0.329	5889	0.35	40%	No	0.344	2047
0.10	40%	Yes	0.333	5722	0.35	40%	Yes	0.348	1853
0.20	0%	No	0.424	690	0.75	0%	No	0.430	335
0.20	0%	Yes	0.436	408	0.75	0%	Yes	0.441	178
0.20	10%	No	0.377	1404	0.75	10%	No	0.375	385
0.20	10%	Yes	0.383	983	0.75	10%	Yes	0.368	244
0.20	20%	No	0.362	2664	0.75	20%	No	0.369	790
0.20	20%	Yes	0.360	2447	0.75	20%	Yes	0.369	619
0.20	30%	No	0.354	2993	0.75	30%	No	0.344	1429
0.20	30%	Yes	0.355	2731	0.75	30%	Yes	0.352	1000
0.20	40%	No	0.333	4424	0.75	40%	No	0.348	2129
0.20	40%	Yes	0.337	4371	0.75	40%	Yes	0.355	1541

3.4. Results and Discussion

In these videos, 1,500,924 faces were recognized, and 326,595 of these faces belonged to children with the rest belonging to the therapists. Table 3.2 shows the emotion distribution of children faces. Some expressions, such as disgust and fear, are recognized with minimal frequency. To explore the usefulness of the extracted features, we have computed the correlations between summarizing features and the CPTI expert annotations. 3.3 shows the top features for each CPTI class.

Since play therapy involves a child playing in a room environment, static cameras are not very successful in capturing continuous streams of face images. The main challenges are the frequent occlusions in the static cameras, resulting in relatively few clear face shots of the child during play, as well as the low resolution of the faces. In our dataset, a child's face is visible only five percent of the time on the average. The distribution of availability is skewed as a power law distribution, and peaks at less than fifty percent for the best sessions. We have used the entire dataset to compute correlations. To prevent overlearning, we have not performed any domain-specific training with the play therapy data itself, but used pre-trained models (i.e., AffectNet) only.

Table 3.2. Distribution of detected emotional expressions.

Emotion	# of Faces
Neutral	95,414
Happy	107,153
Sad	80,915
Surprise	5,298
Anger	830
Fear	168
Disgust	1
None	35,416
Uncertain	1,400

Table 3.3. CPTI vs Framework's Output Correlation

Aggressiveness/Anger	
0.31	Valence-Range
0.25	Sadness
0.25	Arousal-Range
0.24	Arousal-Max100
0.23	Arousal-Max50

Pleasure	
0.21	Valence-Max
0.16	Valence-Max10
0.16	Valence-Max50
0.16	Valence-Max100
0.15	Valence-Max10%

Anxiety/Worry/Wariness	
0.25	Arousal-Mad
0.24	Arousal-Std
0.23	Arousal-Var
0.22	None
0.2	None%

Sadness	
0.32	Anger%
0.3	Anger
0.29	Arousal-Mean
0.29	Arousal-Max100
0.29	Arousal-Max25%

Fear	
0.2	Anger%
0.19	Uncertain
0.18	Uncertain%
0.18	Sadness
0.18	Anger

Shame	
0.28	Uncertain
0.21	Anger
0.14	Neutral%
0.13	Neutral
0.13	Uncertain%

Boredom/Indifference	
0.24	Anger%
0.17	Disgust%
0.17	Disgust
0.14	Neutral%
0.12	Anger

Guilt	
0.29	Arousal-Mode
0.21	Arousal-Min10%
0.21	Arousal-Min25%
0.2	Arousal-Mean
0.2	Arousal-Min

A quick investigation of these values shows that we obtain small to moderate correlations with expert annotations, just by looking at the faces of the children during these sessions. Considering the small percentage of visible faces during the sessions, these results (Table 3.3) are very encouraging.

Our findings in 3.3 indicate a small association between pleasure values as annotated by the experts and our valence values. However, we were not able to find a significant association between pleasure and our basic emotion classes. One possible reason for this could be because pleasure is annotated by CPTI experts under many conditions and is not particular to one basic emotion. Pleasure is scored on the CPTI when there is a reference to content involving happiness, pleasure, satisfaction, and general preference statements with increasing intensity when there is current affective experiencing or activity involving happiness and pleasure.

The small association between anger and arousal-max may underlie the affect regulation difficulties of children with internalizing and externalizing problems who tend to get aroused easily in the face of negative affect, particularly aggression [56]. Proposed that individuals who are low in regulation and high in emotional arousal/intensity are prone to overt expressions of anger and aggressive behaviors. The moderate association between anger as annotated by the CPTI and valence-range could also be another indication of dysregulated emotional intensity, where children fluctuate between intense negative and positive emotions within the same play unit. The small to moderate association between CPTI anxiety and arousal-var could be another indication of affect regulation deficits, indicating that children show significant fluctuations in their arousal levels in the context of anxiety.

We have found small to moderate associations between CPTI sadness, shame, and fear and automatically detected facial anger. This finding indicates that even though these children may show overt signs of anger, the CPTI values as scored by clinicians point to underlying dysphoric affect in the same context. The disruptive and aggressive behaviors of children with externalizing problems have recently been conceptualized as resulting from deficits in affect regulation, which limit the children's ability to cope

with more painful emotions adaptively. These children may use aggressive symptoms to protect themselves from the dysregulation that comes from experiencing dysphoric feelings, such as sadness, shame, and fear [84]. Even though these emotions may not be overtly visible in children's facial cues, the themes that are played out have a dysphoric undertone.

The small to medium associations between guilt and arousal-min is consistent with the conceptualization of guilt as one of the core dysphoric feelings involving self-blame, regret, pangs of conscience, rumination, and sadness [85]. Feelings of guilt arise in response to fears of harming others and reparative concerns, which motivate children to turn inwards, possibly showing less intense overt facial cues in these instances as they try to work out in their internal world reparation of the harm done and restoration of the balance in interpersonal relationships [86].

4. FINE AFFECT ANALYSIS AND RESULTS

In this approach, we focus on creating an automatic affect analysis framework that classifies plays according to CPTI scores. We use both therapist and child data for the therapy session analysis. Moreover, we train the affect analysis system using face modality as well as text modality. Our framework not only investigates single modality performance but also fuses modalities and shows improvements in the results of predicting CPTI. Detailed pipeline can be seen in Figure 4.1.

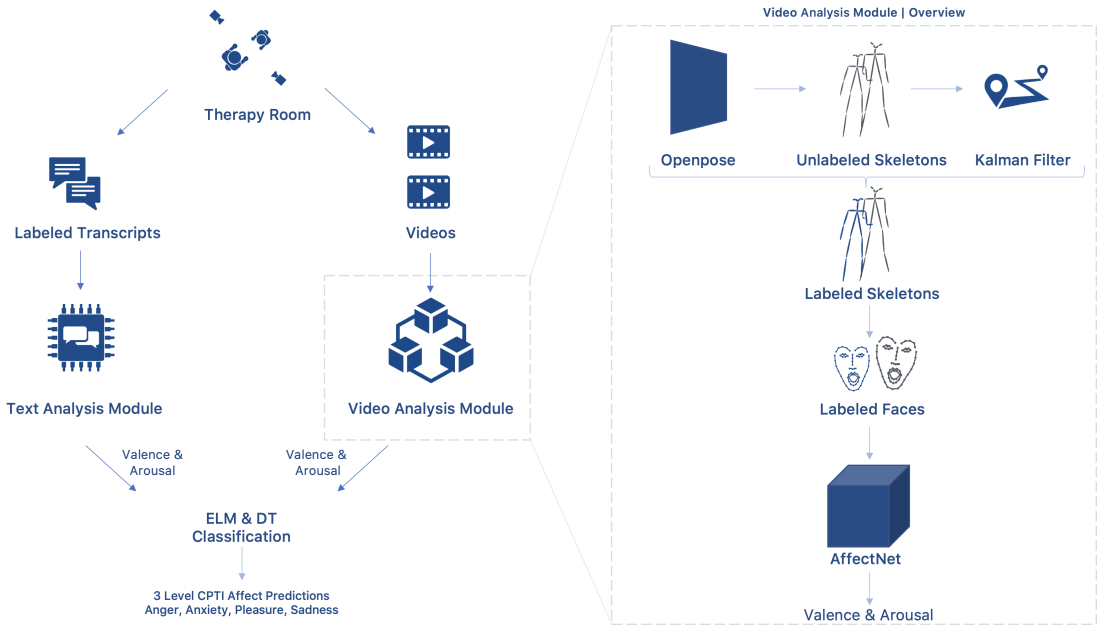

Figure 4.1. Pipeline of fine affect analysis method.

4.1. Tracking

In therapy videos, it is crucial to identify child and therapist to be able to know which emotion belongs to whom. We overcome this problem with face recognition in Chapter 3. However, it is not feasible for automatic analysis because when a new child or/and therapist added to the dataset, the recognizer network has to be retrained. Furthermore, the training process requires time and special hardware such as GPU. For this reason, we focus on tracking the person during the video instead of recognizing s/he in every frame. Besides, recognition is computationally more expensive than

tracking. Therefore, we develop the human tracking module for OpenPose. For tracking we use K-means clustering to cluster detected human skeletons in different groups. Furthermore, we use Kalman filter to correct the tracking skeletons' trajectories. Brief information about K-means clustering and Kalman filter can be found in Sections 4.1.1, and 4.1.2 respectively.

4.1.1. K-means Clustering

K-means is one of the fundamental algorithms in learning. K-means has k centroids where the name of the algorithms comes from. The point belongs to the closest cluster and the algorithm assign point to cluster according to distance with respect to Equation 4.1 and updating centroids according to mean of the belonged points using Equation 4.2.

$$c^{(i)} := argmin_j \left\| x^{(i)} - \mu_j \right\|^2 \tag{4.1}$$

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}} \tag{4.2}$$

Figure 4.2 demonstrates the main logic of the K-means algorithm, and cluster centroids are marked with crosses. In part (a), the plotted dataset is shown without any markup, and in part (b) randomly initialized cluster centroids are marked. Between (c) and (f) two iterations of the algorithm are shown and the algorithm assigns each point to the closest centroid then, cluster centroid is moved to the mean of the assigned points.

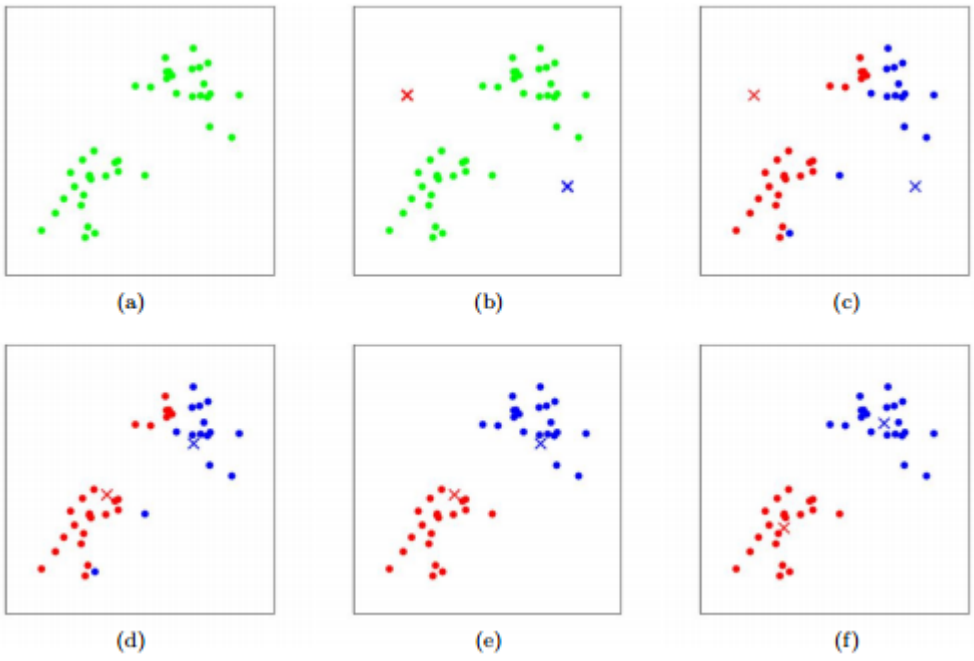

Figure 4.2. K-Means algorithm steps' visualization [87]

4.1.2. Kalman Filter

Kalman Filter (KF) [88] is an estimation algorithm that produces the state estimation of variables based on previous measurements. Moreover, the Kalman Filter also makes a prediction of the future state of the variables based on previous knowledge.

The algorithm predicts the observed variable in the prediction step. If there is an available measurement, the estimation is updated according to a weighted average in the updating step. Estimates with higher certainty has more weight. The filter estimates all errors as gaussian. Kalman filter works with a linear system, on the other hand, Extended Kalman Filter (EKF) is based on the idea of Kalman Filter for the non-linear systems.

Prediction Equation 4.3 shows that the new estimate $\hat{x}_k$ is made by the addition of external influences $B_k \overrightarrow{u}_k$ to the old estimate $\hat{x}_{k-1}$. In addition new uncertainty P_k is achieved by adding environmental impacts Q_k to the old uncertainty P_{k-1}.

$$\begin{aligned} \hat{x}_k &= F_k \hat{x}_{k-1} + B_k \overrightarrow{u}_k \\ P_k &= F_k P_{k-1} + F_k^T + Q_k \end{aligned} \tag{4.3}$$

Kalman algorithm assumes that we predict with some uncertainty. Also, if new measurement data is available, it also has some noise. In update Equation 4.3, we combine sensor measurement gaussian distribution ($\overrightarrow{z_k}$,R_k) and previous prediction gaussian distribution ($H_k\hat{x}_k$,$P_k H_k^T$) with respect to Kalman Gain Equation 4.5. Kalman Gain determines the weight of the measurements.

$$\begin{aligned} \hat{x}_k^{'} &= \hat{x}_k + K^{'}(\overrightarrow{z_k} - H_k \hat{x}_k) \\ P_k^{'} &= P_k - K^{'} H_k P_k \end{aligned} \tag{4.4}$$

$$K^{'} = P_k H_k^T (H_k P_k H_k^T + R_k)^{-1} \tag{4.5}$$

4.1.3. Methodology

The purpose of developing the tracking module is to separate the therapist and the child in the therapy videos. We use OpenPose body landmarks for this separation. In first stage, we start by (a) taking 40x40 crops around the center point between the neck and hip point of each body from the RGB (Red Green Blue) image of that frame, (b) converting it to HSV (Hue Saturation Value) image which is more robust to illumination differences, (c) creating histograms of 10 bins for each channel. The concatenation of these histograms are used as body descriptors in a 2-means clustering algorithm (one for child and one for a therapist, respectively). This also eliminates

incorrect body detections of OpenPose. We use the Kalman filter to be able to track the filtered body detections in the therapy video.

We use OpenPose neck and hip landmarks as our noisy measurements to be tracked. For each new frame, the system predicts the neck and hip points of the detected bodies separately. Then by matching these predicted points to the most likely OpenPose body detections in the new frame, the system identifies each detection as one of the two people. Since Kalman filter has a Gaussian assumption, we use Kalman filters' state covariance matrices as the covariance matrices and the predicted points as the means of the four Gaussian distributions around neck and hip points of child and psychotherapist. After this step, matching the OpenPose body detections to the tracked body locations is achieved by maximizing the joint probability:

$$argmax_{i,j} P(x^i_{neck}, x^i_{hip} | y^j_{neck}, y^j_{hip}) \tag{4.6}$$

where $i \in \{child, therapist\}$, $j \in \{0, 1\}$, x is the our system's prediction and y is the OpenPose body landmark location. Separated OpenPose body landmarks also contain facial landmarks. Therefore, we also separate their faces corresponding to the bodies.

Further examination on the automatic tracking module showed that automatic separation works sufficiently, on the other hand, in some videos, OpenPose generates confusing body landmarks such as there is only one body detection having a child's hip point and psychotherapist's neck point. Also, in some of the videos when the child and the psychotherapist have fast motion while their tracked body landmarks going over one another, the tracking module confuses the bodies. To make our analysis more precise, we manually went over all the videos with 50x speed and the aid of the automatic tracking module to correct any anomalies caused either by OpenPose or by our tracking module. On average, our system tracks humans for 763.2 seconds. In

other words, the tracking system confuses once per every 19,079 frames on the average. After this procedure, we have child and psychotherapist bodies and faces identified separately for each video.

4.2. Feature Extraction

Data is crucial for the learning process because when data size increases, the system makes a better generalization and avoids over-learning. Therefore, we increase our data size using the Horizontal dataset, which is described as detailed in 2.5. Although the data we use is difficult to obtain and challenging to process, we increased the number of children we have from 10 (in Chapter 3) to 52. Also, while expanding our dataset, we paid attention to there are no missing videos, CPTI labels, and transcripts.

After labeling all the OpenPose skeletons with the tracking module, we use the facial landmarks of labeled skeletons to extract facial emotion. For extracting emotional expressions, a similar approach in Chapter 3 is followed; however, differently, we also extract the facial emotion of the therapists. Two pre-trained networks are used to extract facial features, which are emotion classes and continuous valence and arousal.

All plays have transcribed text, and we use a rule-based natural language processing tool that explained in detail in Section 4.2.1 to get valence and arousal scores. The module gives us an opportunity to extracting the valence and arousal scores of the whole sentence or 150-word chunks. Halfon *et al.* developed the tool and used the same problem as a 150-word chunk version to analyze the therapy [10] so we also apply the same method for both therapist and child transcripts.

4.2.1. Text Analysis

The children's native language in our play therapy is Turkish. The entire conversations in plays are transcribed, and since they are in Turkish, the text analysis should be performed with the same language [10]. Keyword spotting is the most obvious approach that matches the words with the pre-defined affect value dictionary.

In a look-up based approach, complex and non-standard sentences are limiting the performance of the system [89]. On the other hand, creating wealthy lexicon is costly and the sentences contain affective words in a very limited portion.

For the Turkish language, there are several natural language processing (NLP) tools for affect analysis [90, 91]. Turkish is an agglutinative language and hundreds of words can be generated from a single root by affixation. Hence, developing a dictionary-based system is quite challenging. Moreover, for Turkish, there is not a widely-used affect dictionary with valence, arousal, and dominance (VAD). SentiTurkNet is a significant work and it provides positivity, negativity, and objectivity scores for each synset in the Turkish WordNet [92].

In this thesis, we use a lexical resource that contains 15,383 words and phrases with their Valence and Arousal affect scores annotated on a five point continuous scale (1-5). We have used a tool developed by Aydın *et al.* to analyze affect in the therapy transcripts and this model model is translated from English lemmas automatically [90] and domain-specific redundant words were eliminated [10]. In addition, the list that has 72 Turkish adverbials, adjectivals, and nominals and 50 interjections are used to boost the analysis. Words in that list can change the affect dramatically. Each word in the dictionary has part of speech (POS) tagging information which was manually tagged by two linguists and these tags are used for calculating the affect score of the sentence. Text analysis can be performed not only in the document level but also in the sentence level. Some words which are used frequently in the therapy sessions have a dramatic affect on the results; therefore, words such as 'father' or 'mother' are being removed. The text analysis tool is developed with Python 2, but our infrastructure works with Python 3. We ported it and also added an API to text analysis tool to retrieve the valence and arousal scores of queried words. In order to classify positive and negative emotions in the Turkish language, converting large corpus from English is more accurate than using reliable and smaller Turkish corpus [90].

Table 4.1. CPTI score distribution among classes

CPTI Score		Class
S >= 4	2	High
4 >S >= 2	1	Medium
S <2	0	Low

4.3. Training

We want to summarize the entire session in a single affect score for a single modality since we have two videos for each session from two cameras and these videos contain a different count of the face. Furthermore, there is the various length of text chunks due to the different amount of conversation during the therapy. Therefore, we get the mean of facial affect scores, as well as text affect scores for both child and therapist separately.

For each method, assume we have a training set of samples, and target labels. When the labels are continuous, such as valence and arousal annotations, the problem is a regression problem, where we seek to map the measured features to a continuous value. When the labels are discrete, such as "internalizing" vs. "externalizing", then the problem is a classification problem. Most machine learning methods can handle both, by some modifications of the mathematical representation. In this work, we use machine learning methods, two of which are decision tree (DT) and extreme learning machines (ELM). Detailed explanation about these methods can be seen in Sections 4.3.1, and 4.3.1.

We express the prediction of CPTI scores as a classification problem into three classes as low, medium, and high, instead of training a regression algorithm for predicting the scores directly. The division can be seen in Table 4.1.

To be able to create a more robust model, combining multiple modalities is a useful approach [93]. It is possible to either combine modalities at the feature level to train a single classifier or at decision (or score) level. We fuse the two modalities

at the feature level since both of their feature spaces consist of valence and arousal dimensions.

We divided the data into three sets as development, training, and test. We used the development set for parameter selection, such as the depth of the decision trees. The development set contains 11 children that show a range of affect scores, according to CPTI. Since we have relatively fewer data points, we use a leave-one-out cross-validation approach for evaluation. In other words, we use 51 children for training and tested with one child from the test set, excluded from training. This approach is applied to every child in the test set, and the mean accuracy is reported.

We follow a different approach for predicting the diagnosis of a child. The diagnosis can belong to four different classes, which are internalizing, externalizing, comorbid, and no diagnosis (i.e., child have no problem) according to the Child Behavior Checklist (CBCL). In other words, the child has a single diagnosis, and we asses the diagnosis class to every play of a child. We use a leave-one-out, cross-validation approach for problem evaluation. We use majority voting among sessions of a single child and assign a single label for diagnosis. If there is an equal distribution for diagnosis prediction, our system randomly selects the diagnosis among the predicted ones.

4.3.1. Decision Tree

A decision tree is a tree-shaped structure and uses a flowchart-based idea. In the tree, nodes represent decisions, and branches represent the outcome of the decisions [94]. The results of decision trees are easy to explain, and it works well even with little data. On the other hand, small changes in the data can have a big effect on the structure of the decision tree. In addition, the depth of the decision tree should be limited to avoid over-learning. Figure 4.3 shows how the information is translated from the table to the decision-tree.

For constructing decision tree information-gain and entropy are used. Entropy is used to calculate the homogeneity of a sample. If the sample is completely homo-

Figure 4.3. Decision tree representation of the table containing information about whether to play golf [95].

geneous, then the entropy is zero, and if the sample is equally divided, then it has an entropy of one. Equation 4.7 and Equation 4.8 stand for entropy calculation and example calculation can be shown in Figure 4.4 and Figure 4.5.

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i \tag{4.7}$$

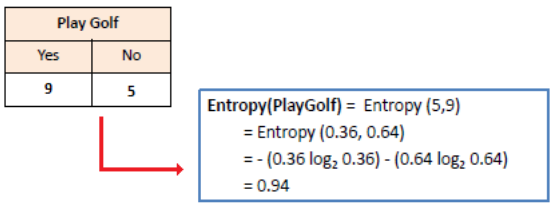

Figure 4.4. Entropy calculation of the frequency table of one attribute [95].

$$E(T, X) = \sum_{c \epsilon X} P(c) E(c) \tag{4.8}$$

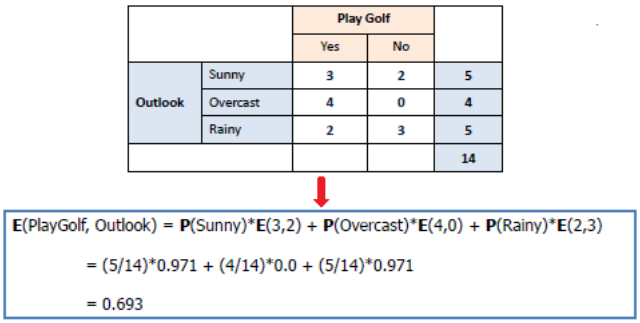

Figure 4.5. Entropy calculation of the frequency table of two attributes [95].

The idea of information gain is to decrease in entropy. For constructing a decision tree, we aim to find an attribute that returns the highest information gain. In other words, we want to get the most homogeneous branches.

$$Gain(T, X) = Entropy(T) - Entropy(T, X) \tag{4.9}$$

In the first step, we calculate the entropy of the target. In the second step, the entropy of each branch is calculated according to information gain Formula 4.9 to find a branch that decreases entropy most. In the end, an attribute has more information gains should be chosen and split dataset on the selected attribute, and this step repeated until reaching a branch with zero entropy.

4.3.2. Extreme Learning Machines

Extreme learning machine (ELM) is a type of feedforward neural networks and used for classification and regression [96]. ELM can have single or multiple layers of hidden nodes, and the weights of these hidden nodes can be randomly assigned, and no update needed for weights such as backpropagation. It also can be used as single-hidden-layer feedforward networks (SLFN). The ELM can be trained very quickly as it does not require a backpropagation to update weights. To train ELM we should

solve the objective function in Equation 4.10, where $h(x)$ denotes features of the D dimensional input and β denotes weight. In Equation 4.11 H specify hidden layer output matrix and T shows the training data target.

$$min_{\beta}||H\beta - T||^2 and||\beta|| \tag{4.10}$$

$$H = \begin{bmatrix} h(x_1) \\ \dots \\ h(x_n) \end{bmatrix} = \begin{bmatrix} h_1(x_1) & \dots & h_L(x_1) \\ \vdots & \ddots & \vdots \\ h_1(x_N) & \dots & h_L(x_N) \end{bmatrix} \tag{4.11}$$

4.4. Results and Discussion

Face detection during therapy for two different sessions can be seen in Figure 4.6 and 4.7. In these figures, the size of the markers coincides with the frequency of detected faces. Figure 4.6 has very few data points from both the therapist and the child. Also, Figure 4.7 has relatively more data points; however, it is also not enough, and the detection rate of the child's face is too low after the half of the session. These figures are good examples to show how difficult the data is and how many missing points we are dealing with.

Table 4.2 shows emotion distribution, and some emotion classes such as fear, contempt, and disgust have a very small frequency. The table shows that the detected therapist face three times higher than the face of the child. The difference can be explained by the fact that the child is more active during the play than the therapist and exhibit an extensive range of body poses.

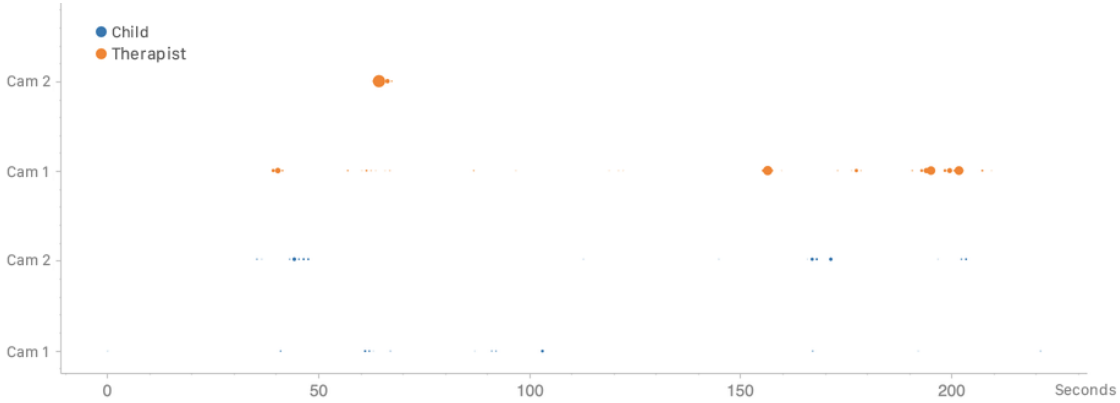

Figure 4.6. Face detection of child and therapist for UAT.

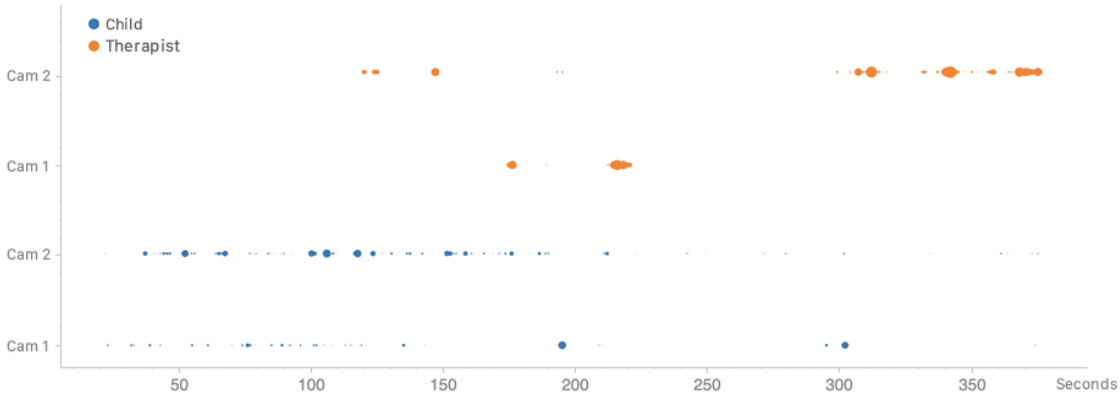

Figure 4.7. Face detection of child and therapist for YZY.

Table 4.3 shows the mean and standard deviation (values in parentheses) of the different modality predictions and CPTI scores on the test set. Although CPTI values between zero and five, we see that we don't actually have that much distributed data, and that the data points are actually very close to each other.

Table 4.4 shows accuracies for three-class CPTI predictions. The random class assignment would result in a 33% accuracy due to three different classes. Prediction of the anger demonstrates the advantage of modality fusion. Using only the child's face to predict anger gives 29% accuracy and only child's text 41%. However, combining these two modalities increase the accuracy of 50%. To predict anxiety, combining a child's modalities gives 47% accuracy; yet using only a therapist's text is the best single-modality result (54%) and combining it with a child's face performs best (57%).

Table 4.2. Distribution of detected emotional expressions in the videos.

Emotion	# of Child Faces	# of Therapist Faces
Neutral	68,910	258,058
Happy	74,323	99,667
Sad	80,691	282,416
Surprise	1,884	3,667
Fear	148	526
Disgust	3	161
Anger	842	30,683
Contempt	0	0
None	33,438	70,062
Uncertain	328	1,553
Total	260,567	746,793

The reason for this could be the therapists are mirroring the child during the therapy, and an automatic system can use this information. In the case of predicting pleasure, the therapist's modalities combined with the child's text modality performs best (74%), but there is a significant point worth mentioning is that as most of the children belong to the normal intensity class, this high accuracy is mostly due to class imbalance. For predicting Sadness, again child's text modality (66%) gives the most information and combining with other modalities does not increase accuracy.

Table 4.4 shows the accuracies of predicting diagnosis classes. We acquire the best performance child facial expression and therapist transcript with 33% accuracy. The performance is slightly above random. The reason behind that could be CBLC checks different aspects of the child such as family conditions, eating habits and school success and there is no direct link between CBCL and the emotions so that it is normal to cause failure to predict diagnosis from CPTI. The distribution of facial affect and text-based affect by diagnosis can be seen in Figure 4.8 and Figure 4.9 respectively. The results show that automatic affect analysis is not accurate enough to provide quantification at a clinically level.

Table 4.3. Mean and Standard Deviation Comparisons of the Different Modality Predictions and CPTI scores on the test set.

Modality	Anger	Anxiety	Pleasure	Sadness
CPTI	1.08 (0.77)	0.76 (0.45)	1.07 (0.28)	0.61 (0.30)
Child Face (CF)	1.17 (0.92)	0.64 (0.29)	0.95 (0.16)	0.72 (0.20)
Child Text (CT)	0.75 (0.91)	0.75 (0.43)	1.03 (0.05)	0.71 (0.21)
Therapist Face (TF)	1.03 (1.01)	0.74 (0.52)	1.05 (0.05)	0.76 (0.18)
Therapist Text (TT)	0.79 (0.94)	0.68 (0.22)	0.96 (0.12)	0.66 (0.23)
CF & CT	0.89 (0.87)	0.78 (0.55)	1.01 (0.04)	0.71 (0.21)
CF & TF	0.80 (0.96)	0.43 (0.44)	0.92 (0.10)	0.80 (0.19)
CF & TT	1.05 (0.72)	0.78 (0.31)	1.03 (0.35)	0.63 (0.24)
CT & TF	0.89 (0.92)	0.63 (0.58)	1.03 (0.08)	0.54 (0.25)
CT & TT	0.75 (0.91)	0.93 (0.46)	0.99 (0.09)	0.68 (0.22)
TF & TT	1.03 (0.93)	1.11 (0.34)	1.04 (0.09)	0.67 (0.28)
CF & CT & TF	0.92 (0.85)	0.64 (0.53)	1.01 (0.04)	0.62 (0.24)
CF & CT & TT	0.84 (0.91)	0.76 (0.34)	1.03 (0.05)	0.71 (0.21)
CF & TF & TT	0.74 (0.89)	1.03 (0.43)	1.04 (0.04)	0.78 (0.20)
CT & TF & TT	0.89 (0.84)	0.71 (0.53)	0.97 (0.08)	0.53 (0.25)
CF & CT & TF & TT	0.92 (0.85)	0.80 (0.48)	1.04 (0.07)	0.62 (0.24)

Table 4.4. Performance Evaluation of the Automated Affect Analysis for CPTI score predictions.

Modality	Anger	Anxiety	Pleasure	Sadness
Child Face (CF)	0.29	0.33	0.59	0.58
Child Text (CT)	0.41	0.49	0.68	0.66
Therapist Face (TF)	0.49	0.42	0.68	0.43
Therapist Text (TT)	0.43	0.54	0.66	0.57
CF & CT	0.50*	0.47	0.68	0.66*
CF & TF	0.47	0.25	0.62	0.41
CF & TT	0.34	0.57*	0.51	0.49
CT & TF	0.42	0.36	0.67	0.49
CT & TT	0.41	0.46	0.72	0.63
TF & TT	0.37	0.36	0.66	0.45
CF & CT & TF	0.36	0.25	0.68	0.61
CF & CT & TT	0.43	0.53	0.70	0.65
CF & TF & TT	0.39	0.28	0.71	0.41
CT & TF & TT	0.37	0.53	0.74*	0.49
CF & CT & TF & TT	0.36	0.42	0.68	0.61
Random	0.33	0.33	0.33	0.33

Table 4.5. Prediction accuracy of diagnosis classes.

Modality	Diagnosis
Child Face (CF)	0.18
Child Text (CT)	0.22
Therapist Face (TF)	0.20
Therapist Text (TT)	0.27
CF & CT	0.21
CF & TF	0.18
CF & TT	0.33*
CT & TF	0.26
CT & TT	0.30
TF & TT	0.27
CF & CT & TF	0.31
CF & CT & TT	0.29
CF & TF & TT	0.30
CT & TF & TT	0.27
CF & CT & TF & TT	0.24
Random	0.25

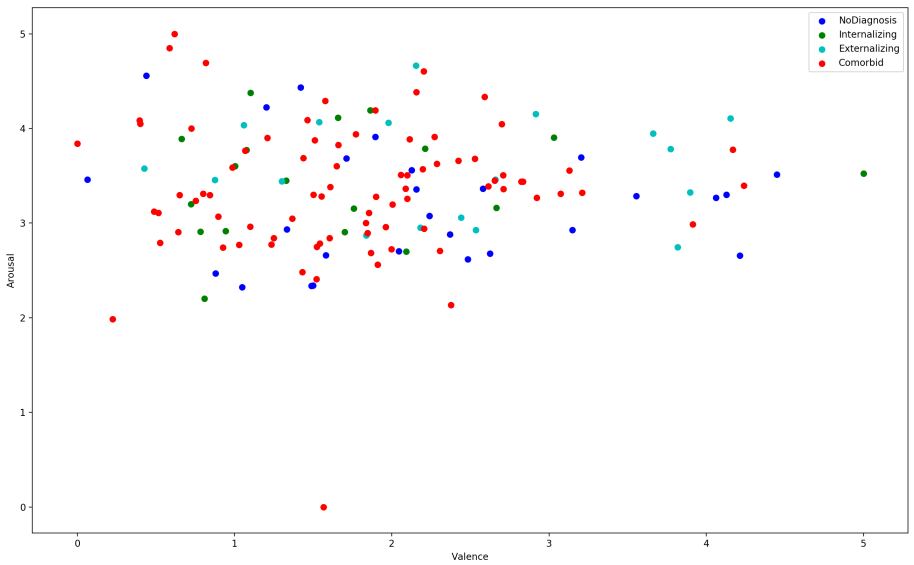

Figure 4.8. Valence and arousal distribution of face for diagnosis classes.

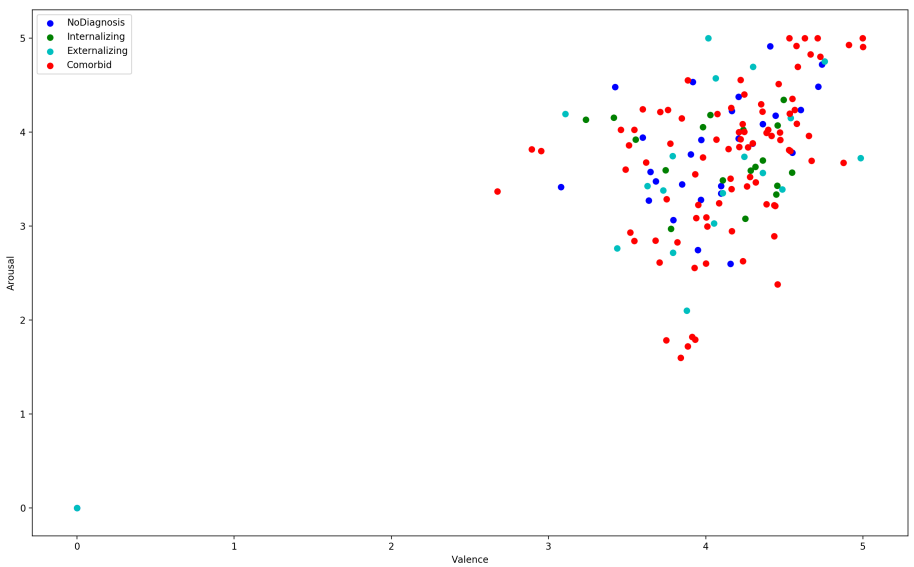

Figure 4.9. Valence and arousal distribution of text for diagnosis classes.

5. FINAL AFFECT ANALYSIS AND RESULTS

With the proposed approach in this section, we predict CPTI affect scores with the regression model. We evaluate both text and face modalities and their fusions. We further investigate the body-movement of the child during therapy and compare how our predictions align with the professional clinicians' assessments. Figure 5.1 shows the proposed pipeline. In this chapter, we also introduce affect analysis tool, which gives an opportunity to the domain experts to investigate the predicted results more deeply.

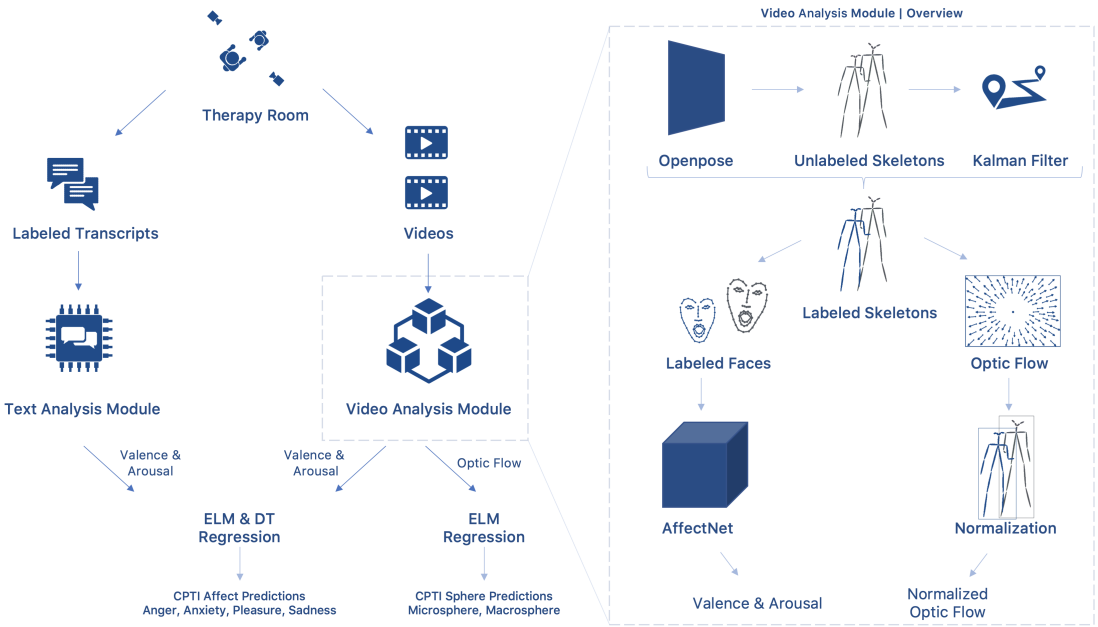

Figure 5.1. Pipeline of proposed affect analysis method.

We can apply regression to be able to predict CPTI scores and get satisfactory results due to the feature selection. In Chapter 4, we get the mean of all valence and arousal scores, which causes high data loss and makes the data meaningless. In other words, getting valence and arousal mean and blend all emotions without considering their emotion classes. This approach has a negative effect on prediction performance and, various feature selection approaches are applied to be able to overcome this problem.

5.1. Feature Extraction and Selection

5.1.1. Facial Feature Selection

We focus on improving the performance of the facial affect analysis and create new data separation, which contains 25 children in the development set. Table 5.1 shows the performance of our method that takes the mean of the whole facial data in the play. Our first approach for feature selection is that we take the means according to circumplex, which has four regions, and the performance can be seen in Table 5.2. Getting the mean by regions improves the anger prediction performance, yet it decrease the performance of the fusion except anger. So far, we have seen that we only use valence and arousal scores produced by AffectNet to get the mean. On the other hand, AffectNet also provides emotion classes, and we decided to average valence and arousal according to emotion classes as our second approach, and the results can be seen in Table 5.3. The emotion-based mean (Table 5.3) boosts fusion performances, and shows the highest accuracy so that we decide to use the mean of every emotion separately.

Table 5.1. Accuracy for three class CPTI by taking overall mean of all data points.

Features	Anger	Anxiety	Pleasure	Sadness
Child Face (CF)	0.29	0.29	0.34	0.34
Therapist Face (TF)	0.36	0.31	0.43	0.17
CF + TF	0.26	0.48	0.43	0.34

Table 5.2. Accuracy for three class CPTI by taking four zones means of all data points.

Features	Anger	Anxiety	Pleasure	Sadness
Child Face (CF)	0.43	0.17	0.43	0.47
Therapist Face (TF)	0.33	0.45	0.28	0.16
CF + TF	0.33	0.38	0.36	0.21

Table 5.3. Accuracy for three class CPTI by taking means of all data points according to the emotion classes.

Features	Anger	Anxiety	Pleasure	Sadness
Child Face (CF)	0.35	0.29	0.38	0.31
Therapist Face (TF)	0.44	0.45	0.36	0.35
CF + TF	0.53	0.51	0.44	0.34

5.1.2. Optic Flow Extraction

CPTI uses microsphere and macrosphere criteria to indicate whether the game takes place in a small space or throughout the room, and we use optical flow to estimate the play area. Optical flow is a vector, and shows the displacement of the objects between two frames. In optical flow, the aim is to reveal the movements of the objects and these movements contain direction and magnitude. The pixels and intensity are the same in the next frame, so we can describe optical flow with the help of intensity I(x, y, t) where $\triangle x$ and $\triangle y$ denotes pixel displacement between two consecutive frames in time $\triangle t$.

$$I(x, y, t) = I(x + \triangle x, y + \triangle y, t + \triangle t) \tag{5.1}$$

Optical flow can be calculated with both sparse and dense techniques. In the sparse approach, only some pixel from the entire image is required. Horn-Schunck [97] and Lucas-Kanade [98] are well-known sparse computation methods for optical flow. Horn-Schunck algorithm tries to minimize distortions in order to achieve smoothness in the flow. The Lucas-Kanade has a different paradigm and assumes constant flow under the local neighborhood. On the other hand, in the dense calculation, all pixels are taken into account. The dense approach is slower but more accurate than sparse due to the consideration of all pixels in the given image. Gunner Farneback's optical flow calculation is an example for dense approach and it uses quadratic polynomials to

compute displacement vector for each neighborhood of the points between two consecutive frames [99]. Both CPU and GPU implementations of Farneback's algorithm are also available in OpenCV library [100].

To extract optical flow, GPU implementation of the OpenCV library is used [100]. Skeleton data from OpenFace is used to create a bounding box area around the child, and we calculated the total scalar magnitude of every point within the bounding box and the total vectors for X and Y directions separately. The optical flow magnitudes change even if the person makes the same movement from the different locations due to the distance so that optical flow should be normalized. Therefore after the optical flow calculation, we normalize it according to the square root of the child's bounding box's area:

$$\frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0 \tag{5.2}$$

$$V_{normalized}(t) = \frac{\sum_{x_{min}}^{x_{max}} \sum_{y_{min}}^{y_{max}} \sqrt{V_x^2 + V_y^2}}{\sqrt{(x_{max} - x_{min})(y_{max} - y_{min})}} \tag{5.3}$$

where x and y denoting pixels in the frame in time t. $\triangle x$,$\triangle y$ and $\triangle t$ indicate the displacement of a point (x, y, t) between two consecutive frames. V_x and V_y specify x and y components of the optical flow. $V_{normalized}(t)$ denotes the normalized optical flow of the child's bounding box at frame t. The width and height of the bounding box are indicated with minimum and maximum pixel indices.

5.2. Training

Due to the application of the emotion based mean and AffectNet gives us labels for 11 emotions, we decide to select meaningful labels to reduce dimensionality. Therefore, we do not use classes that have very few samples such as disgust and contempt; and classes have no information about affect such as uncertain, no-face, and none. We use total five classes which are neutral, happy, sad, fear, anger to train the system and get valence and arousal mean according to these five class so that we have ten feature for each play segment. Moreover, two feature comes from the text analysis tool. For the experiments, we use a development set which has 25 children and applies to leave one child out cross validation for testing. We evaluate our regression results with mean squared error (MSE).

5.3. Affect Analysis Tool

For analyzing verbal and non-verbal affect characteristics of children and therapists during therapy, a tool is required. According to the author' knowledge, there is no such tool developed for play therapy to enable the analysis of longitudinal affect data. For this reason, we develop the tool for domain experts, and this tool helps them to track their affect changes throughout the treatment, and use more detailed session data on demand.

Due to dealing with a large amount of play, our automatic facial and linguistic affect analysis tool uses Elasticsearch for indexing therapy data and Kibana for visualization [101]. Elasticsearch provides full-text search engine with an web interface and Kibana is pluging that works on the Elasticsearch. The affect analysis tool provides interactive visualization on the overall view for affect changes during treatment and filtering for more detailed information, in harmony with Shneiderman's visualization principles [102]. The sessions that have peak data can be seen easily, and the interactive tool allows clinicians to engage related areas and conduct detailed investigations. Our tool has three dashboards. All dashboards support filtering the data with an advanced query structure and exporting them.

The user can see all the children participating in the therapy on the first dashboard. This screen provides the predicted affect scores for each emotion in the treatment mean column, as well as the initial session scores, are shown so that the user can see the affect at the beginning of the therapy. With a date selection feature, the user can select the desired date range and show the children, who get therapy within the selected range. Figure 5.2 shows the first dashboard of the tool.

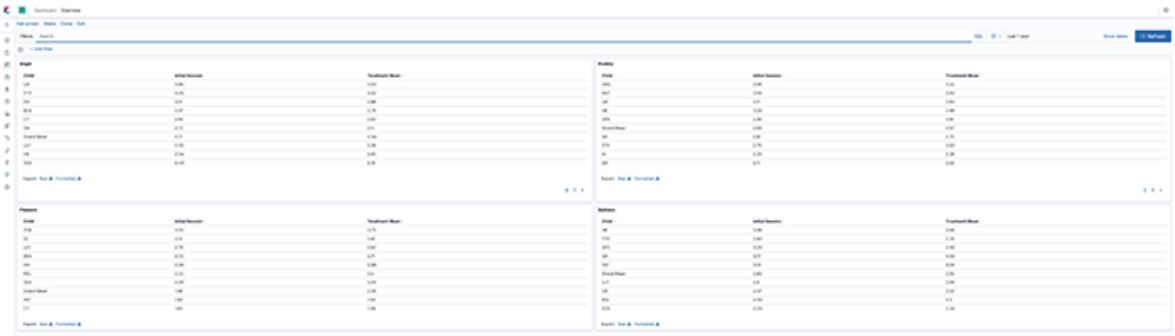

Figure 5.2. Children overview dashboard of the affect analysis tool.

In the second dashboard, the user can see the longitudinal, and session based prediction of the affect data for the selected child. There are also panels that show age, internalizing, and externalizing scores of a child. The user can choose a child using the drop-down menu, and the menu shows all the children have at least one data point in the system. The second dashboard can be seen in Figure 5.3.

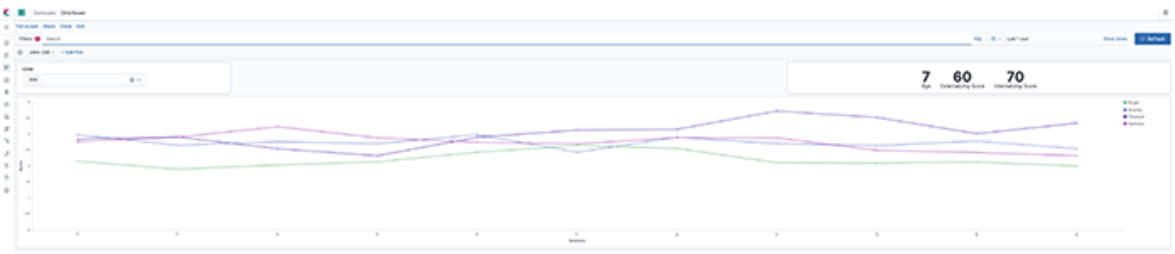

Figure 5.3. Session overview dashboard of the affect analysis tool.

The third dashboard is our last screen, and the purpose of the screen is to investigate play more deeply. Users can see both child's and therapist's longitudinal facial affect. The tool also allows the user to select a child, session, play, and, camera. This dashboard also shows the extreme arousal and valence words as high and low — these words retrieved from the textual analysis module. Figure 5.4 demonstrates the sample

screen.

Figure 5.4. Detailed play overview dashboard of the affect analysis tool.

5.4. Results and Discussion

To ensure the usability of the extracted features, we look at the correlations between the CPTI scores and summarizing function outputs. Table 5.4 shows the highest correlation results for each CPTI category. The combination line shows the weighted average of the two features above them and the weights learned from the training set.

Table 5.4 shows that models we use powerful and can be used to produce valence and arousal scores and CPTI annotations can be predicted with these features. Our following experiments try to predict CPTI with regression methods such as support vector regressors, extreme machine learning regressors, and decision tree regressors.

The performances of different regressors can be seen in Table 5.5, 5.6, 5.7 and 5.8 and leave-one-user-out cross-validation ıs used to measure performance and and random label generator performance is also provided to establish a baseline.

Table 5.4. Framework's output correlation with CPTI scores.

CPTI	Function	Features	Correlation with CPTI
Anxiety	variance	Child Text Arousal	0.35*
		Child Face Arousal	0.10
		Combination	0.35*
Pleasure	maximum	Child Text Valence	0.33
		Child Face Valence	0.33
		Combination	0.40*
Sadness	median	Child Text Arousal	0.44
		Therapist Text Arousal	0.20
		Combination	0.46*
Anger	variance	Child Text Arousal	0.32
		Therapist Text Arousal	0.29
		Combination	0.36*
	minimum	Child Text Valence	-0.26
		Therapist Text Valence	-0.36
		Combination	-0.39*

We analyze the affect regression performances of decision trees (DT), extreme learning machines (ELM), and support vector regressors (SVR) in Table 5.5, 5.6, and 5.7 respectively. First four line of each table shows the performance of the single modality. Modality consist of valence and arousal means. The last line shows the MSE of random generation to create a baseline. Other lines show the fusion performances of these modalities. ELM and SVR perform better than DT for multimodality cases. ELM outperforms SVR and DT in overall results. However, SVR performs better for predicting pleasure. DT shows satisfactory performance in the case of anger and sadness using a child's text, and these results support the finding in Table 5.4. The used pretrained network easily detect happy faces so that it performs satisfying results to find pleasure.

Table 5.5. Mean Square Error between CPTI affect classes and Decision Tree Predictions.

Features	Anger	Anxiety	Pleasure	Sadness
Child Face (CF)	2.92	2.26	1.18	1.75
Child Text (CT)	2.39*	2.19	1.19	1.40*
Therapist Face (TF)	2.73	2.15	1.11*	1.83
Therapist Text (TT)	2.65	1.92	1.14	1.51
CF & CT	2.39	1.88*	1.69	1.40
CF & TF	2.73	2.50	1.18	1.83
CF & TT	3.17	2.15	1.60	1.51
CT & TF	2.85	2.31	1.19	1.40
CT & TT	2.39	2.54	1.38	1.40
TF & TT	2.73	2.15	1.14	1.67
CF & CT & TF	2.91	2.31	1.31	1.40
CF & CT & TT	2.39	2.30	1.47	1.40
CF & TF & TT	2.73	2.15	1.59	1.67
CT & TF & TT	2.88	2.42	1.38	1.40
CF & CT & TF & TT	2.94	2.30	1.31	1.40
Mean Baseline	2.66	1.92	1.13	1.57
Random Generator	5.03	4.60	3.55	4.65

ELM using the fusion of therapist's face and text modalities for predicting anger. Therapists often mirror the child verbal and non-verbal affect states. Especially when the child gets angry, it is hard to understand what s/he says so there is a missing word in the child transcript, however, the therapist rephrase the sayings and explain their actions in a more obvious way, so ELM predicts better scores.

According to results in Table 5.8 optical flow of the horizontal direction and its magnitude represent the microsphere better. On the other hand, the microsphere can be explained better with vertical optical flow. This difference could be explained that moving in the area like the room affects the vertical component of the optical flow because of the getting closer and moving away from cameras. However, playing with

Table 5.6. Mean Squared Error between CPTI affect classes and Extreme Learning Machine Predictions.

Features	Anger	Anxiety	Pleasure	Sadness
Child Face (CF)	2.76	2.03	1.11	1.56
Child Text (CT)	2.50	1.87	1.10	1.50
Therapist Face (TF)	2.64	2.04	1.21	1.60
Therapist Text (TT)	2.59	2.04	1.06	1.63
CF & CT	2.66	2.02	1.11	1.51*
CF & TF	2.77	2.00	1.11	1.65
CF & TT	2.38	1.98	1.02*	1.59
CT & TF	2.71	2.12	1.21	1.57
CT & TT	2.67	2.01	1.14	1.52
TF & TT	2.19*	2.09	1.19	1.69
CF & CT & TF	3.01	1.85*	1.16	1.54
CF & CT & TT	2.74	2.03	1.14	1.54
CF & TF & TT	2.62	1.93	1.18	1.61
CT & TF & TT	3.02	2.04	1.05	1.64
CF & CT & TF & TT	2.81	2.02	1.25	1.52
Mean Baseline	2.66	1.92	1.13	1.57
Random Generator	5.03	4.60	3.55	4.65

miniature toys on the level surface affects the horizontal axis.

Our experiments show that body motion is a good indicator of microsphere and macrosphere predictions, and it is an easy problem for affect prediction. Missing data cause vulnerable facial affect prediction. Therefore, affect prediction accuracy can be increased using a camera located to capture more frontal facial images, and another solution could be dividing play video speech and non-speech segments, which is parallel with transcripts.

Table 5.7. Mean Squared Error between CPTI affect classes and Support Vector Regressor Predictions.

Features	Anger	Anxiety	Pleasure	Sadness
Child Face (CF)	2.80	1.98	1.13	1.72
Child Text (CT)	2.59	1.89	1.13	1.56*
Therapist Face (TF)	2.78	1.88*	1.14	1.66
Therapist Text (TT)	2.67	1.89	1.11	1.71
CF & CT	2.57	2.49	1.09	1.57
CF & TF	2.86	1.90	1.14	1.70
CF & TT	2.70	2.39	1.08	1.76
CT & TF	2.57	1.89	1.17	1.57
CT & TT	2.59	1.89	1.08	1.57
TF & TT	2.61	1.89	1.11	1.64
CF & CT & TF	2.75	1.89	1.14	1.58
CF & CT & TT	2.54	1.89	1.00*	1.57
CF & TF & TT	2.71	2.11	1.37	1.64
CT & TF & TT	2.46*	1.89	1.09	1.58
CF & CT & TF & TT	2.49	1.89	1.07	1.59
Mean Baseline	2.66	1.92	1.13	1.57
Random Generator	5.03	4.60	3.55	4.65

Table 5.8. Mean Squared Error between CPTI movement classes and Decision Tree Predictions.

Features	Microsphere	Macrosphere
X direction (X)	1.70	1.96
Y direction (Y)	1.23	1.84
Total Magnitude (Mag.)	1.19	2.14
X & Y	1.65	1.87
X & Mag.	0.89*	1.76
Y & Mag.	1.00	1.40*
X & Y & Mag.	1.00	1.72
Mean Baseline	1.17	1.78
Random Generator	3.91	3.94

6. CONCLUSION

In this thesis, we examine the contributions of visual and text analysis to therapists' assessments of psychodynamic play therapy sessions with children. Furthermore, we investigate body motion analysis via optical flow to achieve a more detailed analysis of the sessions.

We established a computational tool that operates with a camera-based visual analysis for about a year, which works with the purpose of affect detection of the children who are subjected to such therapy sessions. In doing so, we benefit from the most advanced deep neural networks as well as a set of summarizing functionals to gather information on values over play segments. We have come to the conclusion that automatic analyses are highly correlated with therapists' analyses, even though human experts have greater sources for interpretation.

The automatic system is utilized to estimate CPTI affect dimensions quantitatively and to provide assessments qualitatively. We established that the system can supply significant data, while still being improved towards the better achievement of a more accurate analysis of affect characteristics of children in question.

6.1. Discussion

We apply a setup for realistic recordings in our study in order to form a reliable basis that provides similar outcomes within CPTI. As the text analysis improves, our foundation would improve as well, especially in more in depth affect evaluation. However, the issues we face regarding facial expressions are more of an idiosyncratic nature, thus not subject to much improved expression analysis. Emotions are very difficult to capture by nature. Moreover, we are very reductionist when analyzing emotions, since we have only one label for a long play.

We have established a tool which provides automatic affect analysis through facial expressions and language. We applied state-of-the-art computational tools to play therapy setting in order to help the therapist's assessments of the play therapy sessions. The primary challenge we have faced is regarding capturing the frontal face view with only a few static cameras. The efficiency is open to be improved by adding more resources; however, as a matter of course, this will involve more expenses for the sake of accuracy. It should also be noted that in a modal recording setup, it is not very common to use more than two cameras. The major challenge regarding language-based analysis is the sessions being held in Turkish. However, the tools are rapidly improving.

One of the famous tool Noldus Facereader makes the facial affect analysis, and there are some works using the Facereader to measuring and validating the emotion [103, 104]. There are also cloud services such as Amazon Rekognition, which give user an opportunity to analyze the attributes of faces in images and videos. On the other hand, therapist expertise far beyond the automatic analysis systems, since they can observe more signal than facial and text based affect analysis. Moreover, therapists can shape the course of therapy to get more information and treat the child. However, the purpose of our system is help the therapist to access and interpret therapy more easily. If the therapist can see the change in the data collected for a year and observe anomalies, then automatic affect analysis has a meaning.

There are other considerations about improving the accuracy of the analysis, which concern the paralinguistic analysis of the voices and body motions. Undoubtedly, these will increase the expense as well.

To summarize, the automatic facial analysis tool we have established is hardly sufficient as an only approach, however useful in long-term affect analysis. Moreover, additional improved linguistic expression analysis promotes the accuracy of the system.

The information we expect to receive from various indicators such as valence and arousal, and their range and dynamics is problem-dependent. The modality and the setting of these indicators affect the efficiency of learning approaches when they are

accounted for prediction problems. In order to pursue an accurate outcome, valence and arousal should be sighted together. Thus, the limitation of the system needs to be taken into consideration. Our system is comprehensive and helping the domain experts for exploring the play therapy data.

6.2. Future Work

We applied state-of-the-art methods to play therapy settings and obtained that the main issue regarding face modality is the difficulty of capturing the faces from frontal with two static cameras so that the information on emotions and their valence and arousal degrees would be acquired. Adding more cameras can help improving this condition yet increase the cost, and it is not common that a typical recording setup uses more than two cameras.

It is also possible to run the whole system in near real-time during the therapy and provide a report to the therapist at the end of the session. The results can support the detailed therapy review, and the report can give insights about the session to the expert. On the other hand, working with near real-time learning systems not only requires more data than we process during the work but excessive computation power as well.

Play therapies also have high-quality voice records, and these records can be considered as additional modalities that contribute to improving accuracy. Kaya *et al.* show that audio information helps to provide better emotion recognition accuracy [105].

In this work, we only focus on dimensional emotion modeling, which provides us with valence and arousal scores for face analysis. On the other hand, action units can be used to track the emotional state of the child during the therapy session. Extracting facial action units from face image requires a clear and frontal view of the face. However, it can be used as an additional feature for our pipeline.

Another method may be to make inferences from body movements by using action recognition or considering synchrony between the therapist and the child during the play therapy. Ramseyer *et al.* state that there is little synchrony between the patient and the therapist in the therapy sessions [106, 107]. Our pipeline is suitable to detect the synchrony and affect analysis tool can show synchrony patterns with small enhancement.

REFERENCES

1. Baveye, Y., C. Chamaret, E. Dellandréa *et al.*, "Affective video content analysis: A multidisciplinary insight", *IEEE Transactions on Affective Computing*, Vol. 9, No. 4, pp. 396–409, 2018.

2. Chethik, M., *Techniques of child therapy: Psychodynamic strategies*, Guilford Press, 2003.

3. Zeng, N., H. Zhang, B. Song *et al.*, "Facial expression recognition via learning deep sparse autoencoders", *Neurocomputing*, Vol. 273, pp. 643–649, 2018.

4. Yu, Z. and C. Zhang, "Image based static facial expression recognition with multiple deep network learning", *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 435–442, ACM, 2015.

5. Zhang, Z., P. Luo, C. C. Loy *et al.*, "Facial landmark detection by deep multi-task learning", *European Conference on Computer Vision*, pp. 94–108, Springer, 2014.

6. Wegrzyn, M., M. Vogt, B. Kireclioglu *et al.*, "Mapping the emotional face. How individual face parts contribute to successful emotion recognition", *PloS one*, Vol. 12, No. 5, p. e0177239, 2017.

7. Barrett, L. F., B. Mesquita and M. Gendron, "Context in emotion perception", *Current Directions in Psychological Science*, Vol. 20, No. 5, pp. 286–290, 2011.

8. Ritzi, R. M., D. C. Ray and B. R. Schumann, "Intensive short-term child-centered play therapy and externalizing behaviors in children.", *International Journal of Play Therapy*, Vol. 26, No. 1, p. 33, 2017.

9. Steen, R. L., *Emerging Research in Play Therapy, Child Counseling, and Consultation*, IGI Global, 2017.

10. Halfon, S., E. A. Oktay and A. A. Salah, "Assessing affective dimensions of play in psychodynamic child psychotherapy via text analysis", *HBU*, pp. 15–34, Springer, 2016.

11. Schirmer, A. and R. Adolphs, "Emotion perception from face, voice, and touch: comparisons and convergence", *Trends in Cognitive Sciences*, Vol. 21, No. 3, pp. 216–228, 2017.

12. Poria, S., E. Cambria, R. Bajpai *et al.*, "A review of affective computing: From unimodal analysis to multimodal fusion", *Information Fusion*, Vol. 37, pp. 98–125, 2017.

13. Rudovic, O., Y. Utsumi, J. Lee *et al.*, "CultureNet: A Deep Learning Approach for Engagement Intensity Estimation from Face Images of Children with Autism", *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 339–346, IEEE, 2018.

14. Marinoiu, E., M. Zanfir, V. Olaru *et al.*, "3D Human Sensing, Action and Emotion Recognition in Robot Assisted Therapy of Children with Autism", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2158–2167, 2018.

15. Dibeklioğlu, H., M. O. Hortas, I. Kosunen *et al.*, "Design and implementation of an affect-responsive interactive photo frame", *Journal on Multimodal User Interfaces*, Vol. 4, No. 2, pp. 81–95, 2011.

16. Guha, T., Z. Yang, R. B. Grossman *et al.*, "A computational study of expressive facial dynamics in children with autism", *IEEE transactions on affective computing*, Vol. 9, No. 1, pp. 14–20, 2018.

17. Gongola, J., N. Scurich and J. A. Quas, "Detecting deception in children: A meta-analysis", *Law and human behavior*, Vol. 41, No. 1, pp. 44–54, 2017.

18. Chong, E., K. Chanda, Z. Ye *et al.*, "Detecting Gaze Towards Eyes in Natural Social Interactions and Its Use in Child Assessment", *Proc. ACM IMWUT*, Vol. 1, No. 3, p. 43, 2017.

19. Khan, R. A., A. Crenn, A. Meyer *et al.*, "A novel database of children's spontaneous facial expressions (LIRIS-CSE)", *Image and Vision Computing*, 2019.

20. Batliner, A., S. Steidl and E. Nöth, "Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus", *Proc. of a Satellite Workshop of LREC*, Vol. 2008, p. 28, 2008.

21. Schuller, B., S. Steidl and A. Batliner, "The interspeech 2009 emotion challenge", *Tenth Annual Conference of the International Speech Communication Association*, 2009.

22. Lyakso, E., O. Frolova, E. Dmitrieva *et al.*, "EmoChildRu: emotional child Russian speech corpus", *International Conference on Speech and Computer*, pp. 144–152, Springer, 2015.

23. Kaya, H. and A. A. Karpov, "Efficient and effective strategies for cross-corpus acoustic emotion recognition", *Neurocomputing*, Vol. 275, pp. 1028–1034, 2018.

24. Kaya, H., A. A. Salah, S. F. Gürgen *et al.*, "Protocol and baseline for experiments on Bogazici University Turkish emotional speech corpus", *2014 22nd Signal Processing and Communications Applications Conference (SIU)*, pp. 1698–1701, IEEE, 2014.

25. Monier, F. and S. Droit-Volet, "Synchrony and emotion in children and adults", *International Journal of Psychology*, Vol. 53, No. 3, pp. 184–193, 2018.

26. Larochette, A.-C., C. T. Chambers and K. D. Craig, "Genuine, suppressed and faked facial expressions of pain in children", *Pain*, Vol. 126, No. 1-3, pp. 64–71, 2006.

27. Zeinstra, G. G., M. Koelen, D. Colindres *et al.*, "Facial expressions in school-aged children are a good indicator of 'dislikes', but not of 'likes"', *Food Quality and Preference*, Vol. 20, No. 8, pp. 620–624, 2009.

28. Ekman, P., W. V. Friesen and J. C. Hager, "Facial action coding system: The manual on CD ROM", *A Human Face, Salt Lake City*, pp. 77–254, 2002.

29. Tian, Y.-I., T. Kanade and J. F. Cohn, "Recognizing action units for facial expression analysis", *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 23, No. 2, pp. 97–115, 2001.

30. Littlewort, G. C., M. S. Bartlett, L. P. Salamanca *et al.*, "Automated measurement of children's facial expressions during problem solving tasks", *Face and Gesture 2011*, pp. 30–35, IEEE, 2011.

31. Baltrusaitis, T., A. Zadeh, Y. C. Lim *et al.*, "Openface 2.0: Facial behavior analysis toolkit", *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66, IEEE, 2018.

32. Jaiswal, S. and M. Valstar, "Deep learning the dynamic appearance and shape of facial action units", *2016 IEEE winter conference on applications of computer vision (WACV)*, pp. 1–8, IEEE, 2016.

33. Rudovic, O., J. Lee, M. Dai *et al.*, "Personalized machine learning for robot perception of affect and engagement in autism therapy", *Science Robotics*, Vol. 3, No. 19, p. eaao6760, 2018.

34. Manfredonia, J., A. Bangerter, N. V. Manyakov *et al.*, "Automatic Recognition of Posed Facial Expression of Emotion in Individuals with Autism Spectrum Disorder", *Journal of Autism and Developmental Disorders*, Vol. 49, No. 1, pp. 279–293, Jan 2019, `https://doi.org/10.1007/s10803-018-3757-9`.

35. Joo, H., H. Liu, L. Tan *et al.*, "Panoptic studio: A massively multiview system

for social motion capture", *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3334–3342, 2015.

36. Bennett, C. C. and S. Šabanović, "Deriving minimal features for human-like facial expressions in robotic faces", *International Journal of Social Robotics*, Vol. 6, No. 3, pp. 367–381, 2014.

37. Ekman, P., "Basic emotions", *Handbook of cognition and emotion*, pp. 45–60, 1999.

38. Plutchik, R., "A general psychoevolutionary theory of emotion", *Theories of emotion*, pp. 3–33, Elsevier, 1980.

39. Izard, C. E., "Basic emotions, relations among emotions, and emotion-cognition relations.", American Psychological Association, 1992.

40. Du, S., Y. Tao and A. M. Martinez, "Compound facial expressions of emotion", *Proceedings of the National Academy of Sciences*, Vol. 111, No. 15, pp. E1454–E1462, 2014.

41. Verma, G. K. and U. S. Tiwary, "Affect representation and recognition in 3d continuous valence–arousal–dominance space", *Multimedia Tools and Applications*, Vol. 76, No. 2, pp. 2159–2183, 2017.

42. Wundt, W., "Outlines of psychology", *Wilhelm Wundt and the Making of a Scientific Psychology*, pp. 179–195, Springer, 1980.

43. Russell, J. A., "A circumplex model of affect.", *Journal of personality and social psychology*, Vol. 39, No. 6, p. 1161, 1980.

44. Ekman, P., "Facial action coding system", Consultion Psychologists Press, 1977.

45. Friesen, W. V., P. Ekman *et al.*, "EMFACS-7: Emotional facial action coding system", *Unpublished manuscript, University of California at San Francisco*, Vol. 2,

No. 36, p. 1, 1983.

46. Picard, R. W., *Affective computing*, MIT press, 2000.

47. Zeng, Z., M. Pantic, G. I. Roisman *et al.*, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions", *IEEE transactions on pattern analysis and machine intelligence*, Vol. 31, No. 1, pp. 39–58, 2008.

48. Kleinsmith, A. and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey", *IEEE Transactions on Affective Computing*, Vol. 4, No. 1, pp. 15–33, 2012.

49. Calvo, R. A. and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications", *IEEE Transactions on affective computing*, Vol. 1, No. 1, pp. 18–37, 2010.

50. Gunes, H. and M. Pantic, "Automatic, dimensional and continuous emotion recognition", *International Journal of Synthetic Emotions (IJSE)*, Vol. 1, No. 1, pp. 68–99, 2010.

51. Salah, A. A., H. Kaya and F. Gürpınar, "Video-based emotion recognition in the wild", *Multimodal Behavior Analysis in the Wild*, pp. 369–386, Elsevier, 2019.

52. Kernberg, P. F., S. E. Chazan and L. Normandin, "The children's play therapy instrument (CPTI): description, development, and reliability studies", *The Journal of psychotherapy practice and research*, Vol. 7, No. 3, p. 196, 1998.

53. Achenbach, T. M., "Manual for the Child Behavior Checklist/4-18 and 1991 profile", *University of Vermont, Department of Psychiatry*, 1991.

54. Erol, N. and Z. Şimşek, "Handbook of behaviour assessment scales for school age children and youth (CBCL, YSR and TRF)", Ankara: Metnis Publications, 2010.

55. Boylan, K., T. Vaillancourt, M. Boyle *et al.*, "Comorbidity of internalizing disor-

ders in children with oppositional defiant disorder", *European Child & Adolescent Psychiatry*, Vol. 16, No. 8, pp. 484–494, 2007.

56. Eisenberg, N., T. L. Spinrad and N. D. Eggum, "Emotion-related self-regulation and its relation to children's maladjustment", *Annual review of clinical psychology*, Vol. 6, pp. 495–525, 2010.

57. Eisenberg, N., Q. Zhou, T. L. Spinrad *et al.*, "Relations among positive parenting, children's effortful control, and externalizing problems: A three-wave longitudinal study", *Child development*, Vol. 76, No. 5, pp. 1055–1071, 2005.

58. Halfon, S. and P. Bulut, "Mentalization and the growth of symbolic play and affect regulation in psychodynamic therapy for children with behavioral problems", *Psychotherapy Research*, pp. 1–13, 2017.

59. Christian, K. M., S. Russ and E. J. Short, "Pretend play processes and anxiety: Considerations for the play therapist.", *International Journal of Play Therapy*, Vol. 20, No. 4, p. 179, 2011.

60. Goodfellow, I. J., D. Erhan, P. L. Carrier *et al.*, "Challenges in representation learning: A report on three machine learning contests", *International Conference on Neural Information Processing*, pp. 117–124, Springer, 2013.

61. Arthur, C., A. Meyer, S. Bouakaz *et al.*, "A novel database of children's spontaneous facial expressions (LIRIS-CSE)", *Image and Vision Computing*, 2019.

62. Ringeval, F., A. Sonderegger, J. Sauer *et al.*, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions", *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pp. 1–8, IEEE, 2013.

63. Lucey, P., J. F. Cohn, T. Kanade *et al.*, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression",

2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 94–101, IEEE, 2010.

64. Gross, R., I. Matthews, J. Cohn *et al.*, "Multi-pie", *Image and Vision Computing*, Vol. 28, No. 5, pp. 807–813, 2010.

65. Pantic, M., M. Valstar, R. Rademaker *et al.*, "Web-based database for facial expression analysis", *2005 IEEE international conference on multimedia and Expo*, pp. 5–pp, IEEE, 2005.

66. Mavadati, M., P. Sanger and M. H. Mahoor, "Extended disfa dataset: Investigating posed and spontaneous facial expressions", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–8, 2016.

67. Nicolaou, M. A., H. Gunes and M. Pantic, "Audio-visual classification and fusion of spontaneous affective data in likelihood space", *2010 20th International Conference on Pattern Recognition*, pp. 3695–3699, IEEE, 2010.

68. Langner, O., R. Dotsch, G. Bijlstra *et al.*, "Presentation and validation of the Radboud Faces Database", *Cognition and emotion*, Vol. 24, No. 8, pp. 1377–1388, 2010.

69. McDuff, D., R. Kaliouby, T. Senechal *et al.*, "Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 881–888, 2013.

70. Koelstra, S., C. Muhl, M. Soleymani *et al.*, "Deap: A database for emotion analysis; using physiological signals", *IEEE transactions on affective computing*, Vol. 3, No. 1, pp. 18–31, 2011.

71. Kossaifi, J., G. Tzimiropoulos, S. Todorovic *et al.*, "AFEW-VA database for valence and arousal estimation in-the-wild", *Image and Vision Computing*, Vol. 65,

pp. 23–36, 2017.

72. Barsoum, E., C. Zhang, C. Canton Ferrer *et al.*, "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution", *ACM International Conference on Multimodal Interaction (ICMI)*, 2016.

73. Zafeiriou, S., D. Kollias, M. A. Nicolaou *et al.*, "Aff-Wild: Valence and Arousal'In-The-Wild'Challenge", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 34–41, 2017.

74. Chen, W., O. O. Rudovic and R. W. Picard, "Gifgif+: Collecting emotional animated gifs with clustered multi-task learning", *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 410–417, IEEE, 2017.

75. Mollahosseini, A., B. Hasani, M. J. Salvador *et al.*, "Facial expression recognition from world wild web", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 58–65, 2016.

76. Mollahosseini, A., B. Hasani and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild", *arXiv preprint arXiv:1708.03985*, 2017.

77. Verheugt-Pleiter, A. J., J. E. Zevalkink and M. G. Schmeets, *Mentalizing in child therapy: Guidelines for clinical practitioners.*, Karnac Books, 2008.

78. Cao, Z., G. Hidalgo, T. Simon *et al.*, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields", *arXiv preprint arXiv:1812.08008*, 2018.

79. Karpathy, A., *Convolutional Neural Networks for Visual Recognition*, 2018, `http://web.archive.org/web/20190610155431/https://cs231n.github.io/neural-networks-1/`, accessed at June 2019.

80. Simonyan, K. and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:1409.1556*, 2014.

81. Hassan, M., *VGG16 Convolutional Network for Classification and Detection*, 2018, http://web.archive.org/web/20190501042032/https://neurohive.io/en/popular-networks/vgg16/, accessed at June 2019.

82. Solano, A., *Human pose estimation using OpenPose with TensorFlow*, 2017, http://web.archive.org/web/20190611074753/https://arvrjourney.com/human-pose-estimation-using-openpose-with-tensorflow-part-2-e78ab9104fc8?gi=fb67c17b54f4, accessed at June 2019.

83. Cao, Z., T. Simon, S.-E. Wei *et al.*, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", *CVPR*, 2017.

84. Rice, T. R. and L. Hoffman, "Defense mechanisms and implicit emotion regulation: a comparison of a psychodynamic construct with one from contemporary neuroscience", *Journal of the American Psychoanalytic Association*, Vol. 62, No. 4, pp. 693–708, 2014.

85. Eisenberg, N., "Emotion, regulation, and moral development", *Annual review of psychology*, Vol. 51, No. 1, pp. 665–697, 2000.

86. Fontaine, J. R., P. Luyten, P. De Boeck *et al.*, "Untying the Gordian knot of guilt and shame: The structure of guilt and shame reactions based on situation and person variation in Belgium, Hungary, and Peru", *Journal of cross-cultural psychology*, Vol. 37, No. 3, pp. 273–292, 2006.

87. Piech, C., *K Means*, 2013, http://web.archive.org/web/20190524202932/https://stanford.edu/~cpiech/cs221/handouts/kmeans.html, accessed at June 2019.

88. Kalman, R. E., "A new approach to linear filtering and prediction problems",

Journal of basic Engineering, Vol. 82, No. 1, pp. 35–45, 1960.

89. Mohammad, S. M., "Sentiment analysis: Detecting valence, emotions, and other affectual states from text", *Emotion Measurement*, 2016.

90. Aydın Oktay, E., K. Balcı and A. A. Salah, "Automatic assessment of dimensional affective content in Turkish multi-party chat messages", *Proceedings of the International Workshop on Emotion Representations and Modelling for Companion Technologies*, pp. 19–24, ACM, 2015.

91. Oflazer, K. and M. Saraçlar, *Turkish Natural Language Processing*, Springer, 2018.

92. Dehkharghani, R., Y. Saygin, B. Yanikoglu *et al.*, "SentiTurkNet: a Turkish polarity lexicon for sentiment analysis", *Language Resources and Evaluation*, Vol. 50, No. 3, pp. 667–685, 2016.

93. D'mello, S. K. and J. Kory, "A review and meta-analysis of multimodal affect detection systems", *ACM Computing Surveys (CSUR)*, Vol. 47, No. 3, p. 43, 2015.

94. Quinlan, J. R., "Induction of decision trees", *Machine learning*, Vol. 1, No. 1, pp. 81–106, 1986.

95. Jain, R., *Decision Tree. It begins here.*, 2017, `http://web.archive.org/web/20190501134630/https://medium.com/@rishabhjain_22692/decision-trees-it-begins-here-93ff54ef134`, accessed at June 2019.

96. Huang, G.-B., H. Zhou, X. Ding *et al.*, "Extreme learning machine for regression and multiclass classification", *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, Vol. 42, No. 2, pp. 513–529, 2011.

97. Horn, B. K. and B. G. Schunck, "Determining optical flow", *Artificial intelligence*, Vol. 17, No. 1-3, pp. 185–203, 1981.

98. Lucas, B. D., T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision", Vancouver, British Columbia, 1981.

99. Farnebäck, G., "Two-frame motion estimation based on polynomial expansion", *Scandinavian conference on Image analysis*, pp. 363–370, Springer, 2003.

100. Bradski, G., "The OpenCV Library", *Dr. Dobb's Journal of Software Tools*, 2000.

101. Banon, S., *Open Source Search and Analytics - Elasticsearch.*, 2019, `https://www.elastic.co/`, accessed at June 2019.

102. Shneiderman, B., "The eyes have it: A task by data type taxonomy for information visualizations", *The craft of information visualization*, pp. 364–371, Elsevier, 2003.

103. Terzis, V., C. N. Moridis and A. A. Economides, "Measuring instant emotions during a self-assessment test: the use of FaceReader", *Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research*, p. 18, ACM, 2010.

104. Lewinski, P., T. M. den Uyl and C. Butler, "Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader.", *Journal of Neuroscience, Psychology, and Economics*, Vol. 7, No. 4, p. 227, 2014.

105. Kaya, H., F. Gürpınar and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion", *Image and Vision Computing*, Vol. 65, pp. 66–75, 2017.

106. Ramseyer, F. and W. Tschacher, "Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome.", *Journal of consulting and clinical psychology*, Vol. 79, No. 3, p. 284, 2011.

107. Ramseyer, F. and W. Tschacher, "Synchrony in dyadic psychotherapy sessions",

Simultaneity: Temporal structures and observer perspectives, pp. 329–347, World Scientific, 2008.