Video and Text-Based Affect Analysis of Children in Play Therapy

METEHAN DOYRAN, Utrecht University, The Netherlands
BATIKAN TÜRKMEN, Boğaziçi University, Turkey
EDA AYDIN OKTAY, Northeastern University, US
SIBEL HALFON, Bilgi University, Turkey
ALBERT ALI SALAH, Utrecht University, The Netherlands

Play therapy is an approach to psychotherapy where a child is engaging in play activities. Because of the strong affective component of play, it provides a natural setting to analyze feelings and coping strategies of the child. In this paper, we investigate an approach to track the affective state of a child during a play therapy session. We assume a simple, camera-based sensor setup, and describe the challenges of this application scenario. We use fine-tuned 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. We further investigate text-based and body-movement based affect analysis. We 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.

CCS Concepts: • **Social and professional topics → Children**; • **Applied computing → Health care information systems**; • **Computing methodologies →** *Activity recognition and understanding*; Natural language processing.

Additional Key Words and Phrases: Affective computing, play therapy, emotional expression recognition, deep learning, text-based affect recognition

ACM Reference Format:
Metehan Doyran, Batıkan Türkmen, Eda Aydın Oktay, Sibel Halfon, and Albert Ali Salah. 2019. Video and Text-Based Affect Analysis of Children in Play Therapy. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3340555.3353760

1 INTRODUCTION

Emotion is one of the key aspects of communication, and visual cues during an interaction provide important clues for both the actual and perceived affective states [4][1]. Psychodynamic play therapy provides a natural setting within which children with emotional and behavioral problems are encouraged to express core emotions and social signals through symbolic play (including playing with toys, fantasy, and play involving motor activity). In this framework, the therapist draws attention to the play process by listening actively and inviting the child to communicate in play, encouraging the child to express his or her perceptions, feelings and thoughts. The therapist helps the child reflect on the play context and the accompanying emotions by asking questions about the play setting, temporal ordering of

[1] **This is the uncorrected author proof.**

actions, and the details of the characters, their thoughts, feelings and behaviors in terms of mental states. The therapist also interprets the play context with a wondering stance (e.g. "Why do you think the princess did that?") to help the child see the links between emotions about self and others that find reflection in play behaviors and in the therapeutic relationship, with the purpose of bringing feelings, attitudes, assumptions and beliefs into consciousness. All these interventions help the child organize and have a better understanding of their internal emotional world [6].

In this work, we investigate the possibility of automatic affect analysis of children during psychodynamic play therapy sessions through visual analysis of affect combined with text analysis of the spoken interaction. Automated assessment methods are especially important for continuous monitoring, since survey-like assessments cannot be applied to the child weekly. Considering that every child who is undergoing therapy is producing a wealth of visual information, the annotation effort is significantly costly. Our aim is to facilitate the work of psychologists in their evaluation, help in indexing and visualizing archival material, and to minimize the errors that may arise from a single annotator by using an automated system.

Automatic assessment of affective states in play therapy videos is a challenging task due to the large range of body poses, the existence of various play activities, and occlusions with people and objects. But most importantly, children's expressions of emotions are highly idiosyncratic, and exhibit a large variance depending on the context. We seek to quantify valence and arousal dimensions of affect, as well as basic emotional expressions, both from the faces of the interacting child and therapist, as well as from the transcribed text. We also investigate body motion analysis via optical flow for gaining more insight about the play activity.

The traditional evaluation of play therapy sessions is performed with the Children's Play Therapy Instrument (CPTI) [15], which serves as a gold standard to evaluate our automatic approach. We use six classes from CPTI, which are scored between 0 and 5 separately. Four affect classes assess how much the child shows the following emotions in play; anger, anxiety, pleasure, sadness. Additionally, we have two indicators (microsphere and macrosphere) on how the child uses the play space: A high macrosphere score indicates that the child is playing a game using the entire room (i.e., ball game), on the other hand, low levels of macrosphere show that the game takes place in a smaller area, such as a doll house with symbolic play figures. Similarly, a high microsphere score indicates that the child is using the miniature toy world, as opposed to the entire room to construct their play narrative. Both indices are separately assessed in the CPTI. Usually the child either uses a large space or a small space to play, which causes these two scores to be negatively correlated (with a correlation coefficient of -0.82 in ground truth annotations of our dataset).

The standard treatment plan for the children at the clinic involves once a week therapy session of 50-minutes with the child, along with once a month parent sessions. The treatments are open ended in length and are determined based on progress towards goals, life changes and patients' families' decisions. On average, patients receive 40 sessions over a ten-month period. The CPTI is used for the longitudinal monitoring of these children's play characteristics and progress in affective expression.

A database that contains 270 hours of video collected over a year from child patients is used in this work. Our goal is to analyze long-term emotion changes, of the children as well as the play type (i.e. macrosphere and microsphere) via face, body, and language modalities.

This paper is organized as follows. We first describe related work in assessment of affect during play in Section 2. In Section 3, we describe the building blocks of the proposed assessment framework in detail. The dataset and the experimental setup are introduced in Section 4, followed by our experimental results, comparing the automatic system to the CPTI assessment. We round up with a discussion of the strengths and shortcomings of the work in Section 5. Section 6 concludes the paper.

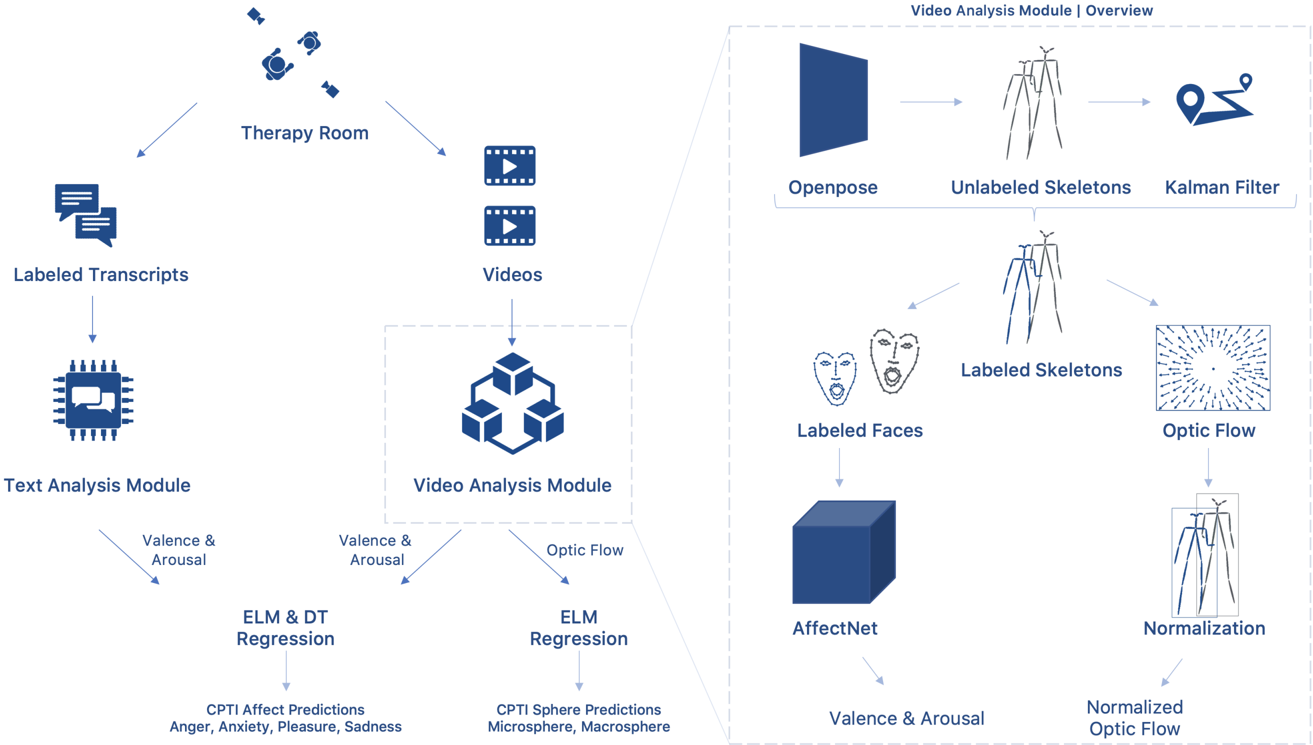

Fig. 1. The schematic layout of the proposed framework.

2 RELATED WORK

Recent developments in affective computing show that machine learning based approaches to affect analysis could be used to achieve models that align quite well with expert judgments [33–35]. Under controlled recording conditions, these models can be trained to infer basic expressions of emotion (happy, sad, angry, fearful, disgusted, surprised) with relatively high reliability, comparable to that of humans [32]. Particularly, deep neural networks, if trained with enough samples (such as AffectNet, trained with 1 million annotated samples [22]) can do a good job of mapping continuous (i.e. valence and arousal) affect of a facial image. Furthermore, it is possible to learn high-level features, such as speed, intensity, irregularity, and extent, to detect emotional expressions in other modalities like speech and gait, and even generalize across modalities [17]. However, there is also evidence that shows that contextual cues strongly shape the interpretation of emotional expressions [2]. Expert humans can take these into account much better than computer-based systems, and are better in detecting subtle changes and rare events. Since it is expensive to train humans and each new annotation requires time and effort, it is worthwhile to understand the capabilities and limits of automatic analysis systems.

Our application setting in this paper is play therapy, which is a common methodology applied for children with emotional and behavioral problems to assist them in reducing their aggressive behaviors. Play is a rich setting for analyzing the actual affective state of children [26, 30]. The relation of affect and play therapy, as well as the assessment of affect during play therapy were reviewed in [11], which proposed a natural language processing-based solution for automatic analysis of affect during play. However, facial and bodily expressions are the most commonly used and researched signals for automatic interaction analysis [28]. Poria et al. presented a detailed overview and showed the advantages of analyzing videos for capturing the affective states, endorsing the reliability of visual data cues and facial expressions [25].

Facial expression analysis for children is a useful tool for settings beyond play. Recently, Guha et al. investigated facial expressions of high functioning autistic children [10]. Such analysis can provide new insights to psychologists, e.g., autistic kids have been found to have less complex dynamics around their eye regions [10]. Some of the analysis tasks are difficult; for instance, detecting deception in children [9]. Automatic observation of face, head pose, and gaze behavior can also reveal analytical insights into social behaviours of children [7]. There is certainly a need for new databases focusing on children's expressions. Khan et al. introduced LIRIS-CSE for children's spontaneous expression recognition [16], but it contains data from 12 children (mean age 7.3), not nearly enough for training classifiers that can generalize well. There is also a clear demand for unobtrusive tools that can deliver timely feedback to the analyst.

We do not analyze speech emotion in this work, but there are computational approaches that specialize in detecting 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 [3]. This dataset, also used for an Interspeech Challenge in 2009, accelerated research in this direction [29]. Later, Lyakso et al. introduced the EmoChildRu corpus, for emotional speech collected from 100 Russian children in the 3–7 year range [18]. We note in passing that it is emotion detection across languages does not work very well, and the training of such systems should be with the same language as the testing, unless elaborate normalization techniques are used [13]. Speech emotion detection resources are very limited for Turkish language [14].

Compared to adults, children show their emotions differently both in terms of speed and variability [23]. Monier et al. hold that this can further change depending on whether or not the child is cooperating with an adult, or even by the level of cooperation. In our application, 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.

It is reasonable to expect that children with special disorders require additional expertise in creating automated tools of analysis for helping psychologists during their therapy procedures [27] or diagnosis processes [19]. Some degree of personalization and tailoring to the special needs of the group seems necessary. In our case, this is achieved by selecting the affect classes according to the measurement needs of the therapy, as well as by adapting the text-based analysis tool by identifying domain-specific words, and their affective values in this particular setting.

3 METHODOLOGY

We assess both video and text based affect analysis in our evaluation framework (see Figure 1). The input to the system consists of two simultaneously recorded video streams from the therapy room, and the transcribed verbal interactions. The ground truth for each play segment is provided by the experts in form of CPTI affect labels per session. We have mainly used correlation analysis for assessing the proposed system. We also used deep learning for measuring valence and arousal. We judge the amount of data to be limited for end-to-end deep learning, but we use pre-trained models for facial affect estimation.

3.1 Video Analysis

We combine and fine-tune several state-of-the-art face analysis components. 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. The video analysis module is shown in Figure 1 (right).

We use the OpenPose tool to detect faces in play therapy videos [5]. OpenPose locates 70 facial landmarks and 25 body landmarks (neck, hip, shoulders, etc.) per subject.

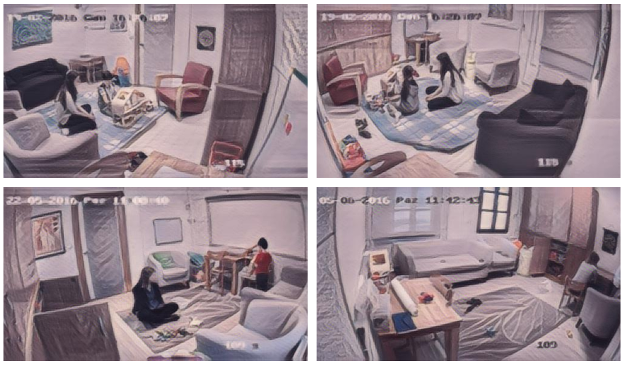

Fig. 2. Sample scenes from the database, processed with a style transfer neural network to preserve the privacy of the participants.

The first task is to reliably identify the therapist and the child in the videos. We use body landmarks for this task. Our approach starts by taking 40x40 crops around the center point between the neck and hip point of each body from a Hue Saturation Value coded image and creates histograms of 10 bins for each channel. The concatenation of these histograms is used as a body descriptor in a 2-means clustering algorithm (for child and therapist, respectively). This also filters out incorrect body detections of OpenPose. The second phase of this module consists of tracking the filtered body detections throughout the video using Kalman filters. The neck and hip landmarks are used as noisy measurements. For each new frame, the system predicts neck and hip points of the detected bodies separately. By matching these predicted points to the most likely OpenPose body detections in the new frame, the system identifies each detection as the child or the therapist. The motion models of the tracked neck and hip points are described by Gaussian distributions. After this step, matching the OpenPose body detections to the tracked body locations is achieved by maximizing the joint probability:

$$argmax_{i,j} P(x^i_{neck}, x^i_{hip} | y^j_{neck}, y^j_{hip}) \tag{1}$$

where $i \in \{child, therapist\}$, $j \in \{0, 1\}$, x is the system's prediction and y is the OpenPose body landmark location.

The automatic identification of the child and therapist works well, but sometimes very fast motion of the child and/or the therapist can create confusions. These are automatically detected, and manually corrected. Our system correctly tracks bodies for 763.2 seconds on average across the database. This means that the system will have a tracking error once per every 19,079 frames. These errors can be automatically detected, as we work in an offline mode.

Once the actors of the interaction are identified, their faces are cropped with a 30% margin with respect to the bounding box of the facial landmarks provided by OpenPose, and frontalized before deep neural network processing. We use two different deep neural networks pre-trained on the AffectNet dataset [22] to determine facial affect. The first neural network we use produces an 11-class output for basic emotional expressions. These classes are neutral, happiness, sadness, surprise, fear, disgust, anger, contempt, none, uncertain, and no-face, respectively. We use it to filter frames to use only the first 7 classes, and to ignore the classes contempt, none, uncertain and no-face. This network also provides a confidence score for each detected emotion. The second deep neural network is used to produce valence and arousal scores for each face. We primarily use these values, in our affect assessment, combined with the confidence scores from the first network.

To characterize the body motion of the child, we have used OpenCV's optical flow implementation, and calculated the total scalar magnitude of every point within the body segment of the child, as well as the total flow in X and Y directions separately. When the children are close to the camera, the optical flow vectors will have a higher magnitude. Thus, a normalization should be applied to eliminate distance effects. After calculating the optical flow, we normalize it according to the square root of the child's bounding box area for each frame:

$$I(x,y,t) = I(x + \triangle x, y + \triangle y, t + \triangle t) \tag{2}$$

$$\frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0 \tag{3}$$

$$V_{normalized}(t) = \frac{\sum_{x_{min}}^{x_{max}} \sum_{y_{min}}^{y_{max}} \sqrt{V_x^2 + V_y^2}}{\sqrt{(x_{max} - x_{min})(y_{max} - y_{min})}} \tag{4}$$

where $I(x,y,t)$ is the intensity image, x and y are the pixel coordinates and t is the frame index of the video. $\triangle x$,$\triangle y$ and $\triangle t$ denote the displacement of a point (x,y,t) between two consecutive frames, computed with the optical flow. V_x and V_y denote x and y components of the optical flow of $I(x,y,t)$. $V_{normalized}(t)$ 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.

The sessions are recorded by two static cameras. Since we have two streams of videos for each session and we would like to summarize the entire session in a single affect score, we use summarizing functionals. For each play segment, the obtained valence and arousal values are passed to the following functionals: mean, minimum, maximum, median, variance. These five summarizing features were extracted for facial valence and arousal of both the child and the therapist.

3.2 Text Analysis

Since we do not have access to high quality voice sampling, we have not attempted a paralinguistic affect analysis on this dataset. However, the entire content of the child's speech is transcribed, and provides a wealth of textual information. Text analysis was performed in Turkish, the children's native language.

The most straightforward approach for automatic affective content analysis of text is based on keyword spotting, which matches a set of detected words to a look-up table that contains keywords and their affective values. The basic limitations of dictionary based approaches are dealing with negation, use of creative and non-standard language and complex sentence structures [21]. Moreover, building a rich lexicon is a very expensive task, and affect-conveying words only form a small portion of a sentence.

Mitigating factors in our case are the complex nature of the language and limited lexical tools and resources. There are only a few natural language processing (NLP) resources in Turkish for affect analysis [1, 24]. Turkish is an agglutinative language, where words can take many suffixes that modify the meaning. In fact, one can generate thousands of word forms from a single root word by affixing with several morphemes. This makes it a lot more challenging to build a dictionary-based system. There is no comprehensive and widely-used lexicon of affective words with valence, arousal, dominance (VAD) annotations for Turkish. A notable work is SentiTurkNet, which provides three polarity scores for each synset in the Turkish WordNet, indicating its positivity, negativity, and objectivity (neutrality) levels [8].

Our textual affect analysis model is based on a semi automatically prepared resource [1] that used automatic translation on a dictionary of English lemmas [31], followed by manual elimination of domain-specific redundant words [11]. Our lexical resource contains 15,383 words and phrases in base form along with their Valence and Arousal affect scores annotated on a five point continuous scale (1-5). These are complemented by a list of 72 Turkish words (adverbials, adjectivals, and nominals) that can intensify or diminish the affective attribute of a sentence and a list of 50 interjections. Apart from the affect scores, all dictionary words include POS tagging information (manually tagged by two linguists). This information is later integrated in some linguistic rules to calculate the overall affect score of a sentence. Additionally, the model handles negation to some extent by detecting both negating words and morphological negating markers. The model is able to calculate a document level (e.g. a single session) affect score, as well as a sentence level affect score (as valence and arousal). Our domain adaptation to play therapy records removes for instance words with normally positive valence (i.e. 'father' or 'mother') that occur frequently in the play therapy setting (as typically they are just outside the therapy room, waiting for their kids), or words involving the objective description of games.

[1] has previously shown that adapting a large corpus from English automatically is in practice more accurate than using a much smaller (but more reliable) Turkish corpus for the classification of positive and negative emotions. However, larger corpus studies are recently planned to provide more extensive affect annotations in Turkish [24], and we do not claim that we have tested the limits of text-based analysis for this setting.

4 EXPERIMENTAL SETUP

Play therapy video records are used in this work are provided by a medical research center which aims to investigate the psychotherapy process of children between 4-10 years old. These children have no drug abuse, no significant developmental delays, no psychotic symptoms, and no significant risk of suicide attempts. 79% of them have internalizing and/or externalizing behavior problems. All 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 Figure 2. Moreover, to be able to get better sound quality, a sound recorder also records the sounds in the room during the therapy. Conversations in all therapy sessions have been transcribed.

Play therapy session videos are divided into subcategories as 'non-play,' 'pre-play,' 'play' and 'interruption by the experts'. Play segments of the videos are annotated by professional clinicians according to CPTI. A total of 302 play segments from 151 therapy sessions of 54 children is used in this work, around 3 sessions per child spaced over a year. In these videos, 1,007,360 faces were detected, and 260,567 of these faces belonged to children with the rest belonging to the therapists. Table 1 shows the emotion distribution of the faces. Some expressions, such as disgust, fear and contempt, are recognized with very small frequency. These rare events can be highlighted for further qualitative analysis.

Since play therapy involves a child playing in a room environment, static cameras are not very successful in capturing continuous streams of face images. 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.

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
Total	**226,801**	**675,178**

Table 1. Distribution of Detected Emotional Expressions in Videos

We separated the dataset randomly into training and test sessions with no subject overlap. We chose the functionals with the highest correlations and learned the parameters to combine different features together over the training dataset. The test dataset is only used with the best performing functionals and combination parameters chosen from the training experiments. In Table 2 we present only the test results. The facial analysis modules were not fine-tuned with the play therapy data, due to the lack of frame-level ground truth annotations.

In our assessments, we looked at the correlations of our system's valence and arousal predictions with CPTI labels, which are real values between 0 and 5. We have additionally contrasted two regression approaches (using decision trees and extreme learning machines, respectively) to directly predict these labels. For these experiments, we separate a development set from the database, and optimize the meta-parameters of the regressors on the development set. We test these systems with leave-one-user-out cross-validation on the remaining data. We choose our main evaluation metric to be the mean squared error (MSE) for regression.

5 RESULTS AND DISCUSSION

To explore the usefulness of the extracted features, we have computed the correlations between summarizing features and the CPTI expert annotations. Table 2 shows the top features for each CPTI class. For anxiety, pleasure and sadness, we find almost no negative correlations with our valence and arousal features, therefore we excluded negative correlations from these classes. The lines marked as "Combination" represent the weighted average of the two features above that line with the weights learned from the training set.

Our findings in Table 2 indicate a moderate association between children's anxiety values as annotated by the experts and variance of children's text arousal values. Combining it with the face arousal values does not change the correlation. This association between CPTI anxiety and arousal variance could be another indication of affect regulation deficits, indicating that children show significant fluctuations in their arousal levels in the context of anxiety.

There is moderate association between children's pleasure values and maximum of children's text and face valence values. Combination of these two increases the correlation significantly. These findings are aligned with the known relationship of valence values with the pleasure emotions [20]. Maximum of the valence values indicates the child's pleasure intensity.

The strongest single feature association is found between the median of children's text arousal and sadness CPTI scores. Combining it with the therapist's text arousal scores further increases the correlation up to 0.46. Although this correlation may seem contradicting with the known relationship between sadness emotion and valence/arousal space, it can be hinting at some imperfections in our text analyzing system. It is also possible that our system picks up certain signals of agitation in children with a clinical diagnosis, that would not be encountered in children with no

CPTI	Function	Features	Corr. 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**

Table 2. Framework’s output correlation with CPTI scores

diagnosis. Children with internalizing and externalizing problems have difficulty tolerating feelings of sadness [12] and can become overly agitated and anxious when they face feelings of sadness. The increase in arousal may have to do with their affect regulation deficits when they experience sad feelings.

There is moderate association between anger, as annotated by the CPTI and children’s text arousal variance. This could be an indication of dysregulated emotional intensity, where children fluctuate between intense high and low arousal within the same play unit. When children experience angry feelings, especially children with externalizing problems, it is possible that they have difficulty toning down the aggression, and engage in overly arousing play and can only stop themselves fully with the limits set by the therapist, and hence the low arousal scores. Combining it with therapist’s text arousal seems to improve the correlation with the children’s anger CPTI score. A detailed manual investigation over the texts proves that many words are missing or intentionally misspelled from the children’s script, mostly because it is difficult to transcribe it over a microphone record and to keep the original mispronounced words that the children say. Moreover we see that therapists very frequently repeat the mispronounced words or try to guess the mumbling speeches. Therefore including therapist’s scripts allows us to gather more information about the children’s emotions throughout the therapy. Similar findings are achieved through combining children’s text valence with therapist’s text valence scores. Applying minimum function over this combination improves the negative correlation with the anger CPTI score. This negative correlation is also in line with the association between valence scores to anger emotion.

Overall Table 2 suggests that the pre-trained models we use are strong enough to generate meaningful valence and arousal scores which can be used to predict expert annotated CPTI emotion scores. Our second set of experiments tries to achieve this by using well known regression techniques, such as decision tree regressors and extreme machine learning regressors.

In Table 3, Table 4 and Table 5 we present the leave-one-user-out cross validation performances of the decision tree regressors and the extreme learning machine regressors. A random label generator is also given to indicate a baseline.

Features	Anger	Anx.	Pleas.	Sad.
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
Baseline	5.03	4.60	3.55	4.65

Table 3. MSE between CPTI affect classes (Anger, Anxiety, Pleasure, Sadness) and DT predictions

Table 3 and Table 4 show CPTI affect label predictions using textual and facial features with the help of decision trees (DT) and extreme learning machines (ELM), respectively. The first four lines correspond to the mean of the valence and arousal values separately of that modality. The other lines show the feature level combinations of these four modalities. In our experimental setting, ELMs handle multi-modality much better than DTs. The best performing features for ELMs are mostly the combinations of two modalities. On the other hand, the best performing features for DTs are mostly single modalities. The best predictions of DTs are for the anger and sadness CPTI affect classes.These two best DT predictors only use child's text valence and arousal scores, which is in line with the correlation findings in Table 2. There is also improvement in MSE with ELM pleasure predictor using the combination of children's face modality and therapist's text modality. Happiness can be easily found by the pretrained network we use, therefore using child's face modality predicts pleasure better. It appears that the therapists mirror children's pleasure levels.

The best predictor of all the CPTI affect classes is the ELM predictor (Table 4) for the anger affect class, using a combination of therapist's face and text modalities. For valence, this may suggest that the increase in therapist's close verbal and non-verbal mirroring of the child's affect states (especially when the child gets angry) can be utilized by the ELM method we use to predict the child's anger level. In the text modality, we notice that when the child gets angry, it is harder to understand his/her words and therefore there are missing words in his/her sentences from the text, but the therapists rephrase them or explain their actions in a clearer way, thus helping the ELM to extract better information to predict the child's anger level.

Table 5 shows that the optical flow in the horizontal direction and magnitude express microsphere better than macrosphere. However, this situation changes in favor of the vertical direction when we want to explain the microsphere. These results can be explained by movement in larger area increases vertical direction due to the changing distance from camera. On the other hand, when children play with toys on flat surface or lie down, it is more likely to have optical flow changes on the horizontal axis.

Overall, our evaluations show that microsphere and macrosphere predictions via body motion detection is an easier problem compared to affect label predictions, which is plagued by missing data, as well as idiosyncratic variations.

Features	Anger	Anx.	Pleas.	Sad.
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
Baseline	5.03	4.60	3.55	4.65

Table 4. MSE between CPTI affect classes (Anger, Anxiety, Pleasure, Sadness) and ELM predictions

Affect label predictions can possibly be improved by using a setting in which the cameras are positioned to capture more facial images. Another possible point of improvement is to separate the facial changes during speech from the non-speech segments. This requires a good alignment between the multiple cameras and diarized speech signal.

Features	Microsphere	Macrosphere
Horizontal flow (Hor.)	1.70	1.96
Vertical flow (Ver.)	1.23	1.84
Total Magnitude (Mag.)	1.19	2.14
Hor. & Ver.	1.65	1.87
Hor. & Mag.	**0.89**	1.76
Ver. & Mag.	1.00	**1.40**
Hor. & Ver.& Mag.	1.00	1.72
Baseline	3.91	3.94

Table 5. MSE between CPTI play area utilization classes and DT predictions based on optical flow in different directions

6 CONCLUSIONS

We have investigated an application where children's affect is tracked during play therapy sessions via camera-based face analysis for about a year of therapy. We used state-of-the-art deep neural networks in conjunction with a set of summarizing functionals to pool values over play segments, and observed that the automatic assessment shows promising correlations with therapist assessments, even though the therapist has access to a much larger set of signals for interpretation, and the automatic system could only capture a small percentage of faces plus some partially missing scripts. A system based on our approach can provide the therapist with overview visualizations, help with the retrieval of stored video episodes where a particular emotion is prominent, and to assess patient-therapist interaction quality.

Our work provides a good baseline through operating with real data close to in-the-wild settings and a realistic recording setup. This baseline estimates what can be expected for instance through processing of legacy recordings of session data in similar cases. Improvement in text processing, particularly for more elaborate affect assessment, would improve this baseline. The problems with facial expressions, however, appear to be more idiosyncratic, and unlikely to benefit much from improved expression analysis.

To summarize, our results show that automatic facial analysis is useful to some extent in long-term affect monitoring, but it is not sufficient as a single modality. Combining it with linguistic expression analysis improves the system's performance. The informativeness of different indicators (such as valence, arousal, as well as their range and dynamics) is problem-dependent. When these indicators are treated as prediction problems, the success rate of machine learning approaches depend on the modality and the setting. For example, both valence and arousal are relevant indicators that need to be observed jointly for a healthy diagnosis. However in a camera-based setup, arousal is easier to estimate, as both optical flow based body motion and facial analysis could contribute to its estimation. Subsequently, both theoretical and practical issues are relevant in developing automatic solutions. These limitations need to be taken into account when interpreting the outcomes of the automatic systems. Our contribution is a step towards a comprehensive system for helping therapists for indexing and exploring of play therapy data.

REFERENCES

[1] Eda Aydın Oktay, Koray Balcı, and Albert Ali Salah. 2015. Automatic assessment of dimensional affective content in Turkish multi-party chat messages. In *Proc. Int. Workshop on Emotion Representations and Modelling for Companion Technologies*. ACM, 19–24.

[2] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. 2011. Context in emotion perception. *Current Directions in Psychological Science* 20, 5 (2011), 286–290.

[3] A Batliner, S Steidl, and E Nöth. 2008. Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus. In *Workshop on Corpora for Research on Emotion and Affect*.

[4] Yoann Baveye, Christel Chamaret, Emmanuel Dellandréa, and Liming Chen. 2018. Affective video content analysis: A multidisciplinary insight. *IEEE Transactions on Affective Computing* 9, 4 (2018), 396–409.

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

[6] Morton Chethik. 2003. *Techniques of child therapy: Psychodynamic strategies*. Guilford Press.

[7] Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M Jones, Agata Rozga, and James M Rehg. 2017. Detecting Gaze Towards Eyes in Natural Social Interactions and Its Use in Child Assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 43.

[8] Rahim Dehkharghani, Yucel Saygin, Berrin Yanikoglu, and Kemal Oflazer. 2016. SentiTurkNet: a Turkish polarity lexicon for sentiment analysis. *Language Resources and Evaluation* 50, 3 (2016), 667–685.

[9] Jennifer Gongola, Nicholas Scurich, and Jodi A Quas. 2017. Detecting deception in children: A meta-analysis. *Law and human behavior* 41, 1 (2017), 44–54.

[10] Tanaya Guha, Zhaojun Yang, Ruth B Grossman, and Shrikanth S Narayanan. 2018. A computational study of expressive facial dynamics in children with autism. *IEEE Trans. Affective Computing* 9, 1 (2018), 14–20.

[11] Sibel Halfon, Eda Aydın Oktay, and Albert Ali Salah. 2016. Assessing affective dimensions of play in psychodynamic child psychotherapy via text analysis. In *Proc. HBU*. 15–34.

[12] Leon Hoffman, Timothy Rice, and Tracy Prout. 2015. *Manual of regulation-focused psychotherapy for children (RFP-C) with externalizing behaviors: A psychodynamic approach*. Routledge.

[13] Heysem Kaya and Alexey A Karpov. 2018. Efficient and effective strategies for cross-corpus acoustic emotion recognition. *Neurocomputing* 275 (2018), 1028–1034.

[14] Heysem Kaya, Albert Ali Salah, Sadık Fikret Gürgen, and Hazım Ekenel. 2014. Protocol and baseline for experiments on Bogazici University Turkish emotional speech corpus. In *Proc. IEEE SIU*. 1698–1701.

[15] Paulina F Kernberg, Saralea E Chazan, and Lina Normandin. 1998. The children's play therapy instrument (CPTI): description, development, and reliability studies. *The Journal of psychotherapy practice and research* 7, 3 (1998), 196.

[16] Rizwan Ahmed Khan, Arthur Crenn, Alexandre Meyer, and Saida Bouakaz. 2019. A novel database of children's spontaneous facial expressions (LIRIS-CSE). *Image and Vision Computing* (2019).

[17] Angelica Lim and Hiroshi G Okuno. 2014. The mei robot: towards using motherese to develop multimodal emotional intelligence. *IEEE Transactions on Autonomous Mental Development* 6, 2 (2014), 126–138.

[18] Elena Lyakso, Olga Frolova, Evgeniya Dmitrieva, Aleksey Grigorev, Heysem Kaya, Albert Ali Salah, and Alexey Karpov. 2015. EmoChildRu: emotional child Russian speech corpus. In *International Conference on Speech and Computer*. Springer, 144–152.

[19] Joseph Manfredonia, Abigail Bangerter, Nikolay V. Manyakov, Seth Ness, David Lewin, Andrew Skalkin, Matthew Boice, Matthew S. Goodwin, Geraldine Dawson, Robert Hendren, Bennett Leventhal, Frederick Shic, and Gahan Pandina. 2019. Automatic Recognition of Posed Facial Expression of Emotion in Individuals with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders* 49, 1 (2019), 279–293.

[20] Mika Mäntylä, Bram Adams, Giuseppe Destefanis, Daniel Graziotin, and Marco Ortu. 2016. Mining Valence, Arousal, and Dominance: Possibilities for Detecting Burnout and Productivity?. In *Proc. Int. Conf. on Mining Software Repositories*. 247–258.

[21] Saif M Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion Measurement* (2016).

[22] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affective Computing* 10, 1 (2017), 18–31.

[23] Florie Monier and Sylvie Droit-Volet. 2018. Synchrony and emotion in children and adults. *International Journal of Psychology* 53, 3 (2018), 184–193.

[24] Kemal Oflazer and Murat Saraçlar. 2018. *Turkish Natural Language Processing*. Springer.

[25] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.

[26] Rochelle M Ritzi, Dee C Ray, and Brandy R Schumann. 2017. Intensive short-term child-centered play therapy and externalizing behaviors in children. *International Journal of Play Therapy* 26, 1 (2017), 33.

[27] Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W Picard. 2018. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics* 3, 19 (2018), eaao6760.

[28] Annett Schirmer and Ralph Adolphs. 2017. Emotion perception from face, voice, and touch: comparisons and convergence. *Trends in Cognitive Sciences* 21, 3 (2017), 216–228.

[29] Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The Interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*.

[30] Rheta LeAnne Steen. 2017. *Emerging Research in Play Therapy, Child Counseling, and Consultation*. IGI Global.

[31] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* 45, 4 (2013), 1191–1207.

[32] Martin Wegrzyn, Maria Vogt, Berna Kireclioglu, Julia Schneider, and Johanna Kissler. 2017. Mapping the emotional face. How individual face parts contribute to successful emotion recognition. *PloS one* 12, 5 (2017), e0177239.

[33] Zhiding Yu and Cha Zhang. 2015. Image based static facial expression recognition with multiple deep network learning. In *Proc. ACM on International Conference on Multimodal Interaction*. ACM, 435–442.

[34] Nianyin Zeng, Hong Zhang, Baoye Song, Weibo Liu, Yurong Li, and Abdullah M Dobaie. 2018. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing* 273 (2018), 643–649.

[35] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*. Springer, 94–108.