Automatic Behavioral Analysis from Faces and Applications to Risk Marker Quantification for Autism

by

Jordan Hashemi

Department of Electrical and Computer Engineering
Duke University

Date: ____________________________

Approved:

__________________________
Guillermo Sapiro, Supervisor

__________________________
Carlo Tomasi

__________________________
Robert Calderbank

__________________________
Lawrence Carin

__________________________
Vahid Tarokh

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering in the Graduate School of Duke University

2018
ABSTRACT

Automatic Behavioral Analysis from Faces and Applications to Risk Marker Quantification for Autism

by

Jordan Hashemi

Department of Electrical and Computer Engineering
Duke University

Date: ______________________

Approved:

__________________________
Guillermo Sapiro, Supervisor

__________________________
Carlo Tomasi

__________________________
Robert Calderbank

__________________________
Lawrence Carin

__________________________
Vahid Tarokh

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering in the Graduate School of Duke University 2018
Abstract

This dissertation presents novel methods for behavioral analysis with a focus on early risk marker identification for autism. We present current contributions including a method for pose-invariant facial expression recognition, a self-contained mobile application for behavioral analysis, and a framework to calibrate a trained deep model with data synthesis and augmentation. First we focus on pose-invariant facial expression recognition. It is known that 3D features have higher discrimination power than 2D features; however, usually 3D features are not readily available at testing time. For pose-invariant facial expression recognition, we utilize multi-modal features at training and exploit the cross-modal relationship at testing. We extend our pose-invariant facial expression recognition method and present other methods to characterize a multitude of risk behaviors related to risk marker identification for autism. In practice, identification of children with neurodevelopmental disorders requires low specificity screening with questionnaires followed by time-consuming, in-person observational analysis by highly-trained clinicians. To alleviate the time and resource expensive risk identification process, we develop a self-contained, closed-loop, mobile application that records a child’s face while he/she is watching specific, expertly-curated movie stimuli and automatically analyzes the behavioral responses of the child. We validate our methods to those of expert human raters. Using the developed methods, we present findings on group differences for behavioral risk markers for autism and interactions between motivational framing context, facial
affect, and memory outcome. Lastly, we present a framework to use face synthesis to calibrate trained deep models to deployment scenarios that they have not been trained on. Face synthesis involves creating novel realizations of an image of a face and is an effective method that is predominantly employed only at training and in a blind manner (e.g., blindly synthesize as much as possible). We present a framework that optimally select synthesis variations and employs it both during training and at testing, leading to more efficient training and better performance.
Contents

Abstract iv
List of Tables x
List of Figures xi
Acknowledgements xiii

1 Introduction 1
1.1 Cross-modality and pose-invariant facial expression recognition 2
1.2 Computer vision analysis for quantification of autism risk behaviors 2
1.3 Findings and extensions using computer vision analysis 3
1.4 Framework for intelligent synthesis driven model calibration 3
1.5 Organization of the dissertation 4

2 Cross-modality and pose-invariant facial expression recognition 5
2.1 Chapter summary 5
2.2 Introduction and related work 6
2.3 Proposed framework 9
2.3.1 Pose-invariant dictionary 11
2.3.2 Cross modality and domain representation 12
2.4 Experimental validation 13
2.4.1 Datasets and setup 13
2.4.2 Validation on BU3D-FE 15
3 Computer vision analysis for quantification of autism risk behaviors

3.1 Chapter summary

3.2 Introduction

3.3 Data collection

3.3.1 Setup and participants

3.3.2 Movie stimuli

3.4 Methods

3.4.1 Automatic coding

3.4.2 Human coding

3.5 Validation results and analysis

3.5.1 Validation results

3.5.2 Behavioral analysis

3.6 Discussion

3.6.1 Limitations

3.7 Conclusion

4 Findings and extensions using computer vision analysis

4.1 Chapter summary

4.2 Atypical attention in toddlers with autism

4.2.1 Motivation and objectives

4.2.2 Methods

4.2.3 Results

4.2.4 Discussions and conclusions

4.3 Automatic emotion and attention analysis of young children at home
6 Conclusion ............... 99

Bibliography .............. 101

Biography ................. 112
List of Tables

2.1 Comparisons of recognition rates (%) for varying expressions across different methods on BU3D-FE dataset. .......................... 15

2.2 Comparisons of recognition rates for all expressions across multiple poses. ................................................. 17

3.1 Demographics table reported in means (standard deviations) or counts (percentage) .................................................. 25

3.2 Number of video recordings .................................................. 31

3.3 Inter-rater reliability results of automatic codings to expert human raters. Reported are ICC (95% CI) ................................. 33

3.4 Overall validation results of the automatic emotion classification with expert human raters. ................................. 36

3.5 Per group validation results of the automatic emotion classification with expert human raters. ................................. 36

4.1 Number of children in each group orienting to name inconsistently (less than twice) vs consistently (twice or more). ............... 51

4.2 Reported number of movies viewed and whether the videos could be analyzed using our automatic video coding .................. 56

5.1 \textit{OFD} rank-1 classification scores across testing setups for varying training setups and synthesis selection parameter for training. .... 94

5.2 Results on calibrating the \textit{VGG-Face} model to the \textit{OFD} dataset ........... 96

5.3 Results on CASIA NIR-VIS 2.0 benchmark. ................................. 97
# List of Figures

2.1 Overview of multi-modality and pose-invariant dictionary construction. 8

2.2 Overview and example of the proposed cross-modality and pose-invariant representation. 10

2.3 Confusion matrix result for facial expression recognition across all poses. 18

3.1 Examples of the recorded video from the front facing tablet’s camera, of automatic facial landmarking, and screenshots from movie stimuli. 26

3.2 Example of a head turn using automatic method. 29

3.3 Automatic name-call prompt detection. 29

3.4 Area plots of absolute time difference between automatic methods and hand labeled data for name-call prompt and for head turn detection. 35

3.5 Examples of responses from a non-ASD and ASD toddler to name-call. 40

3.6 Examples of the head movement of three participants during the Puppets stimuli. 41

3.7 Probability scores of expressing Happy for a non-ASD and ASD toddler reacting to a scene during the Bunny movie stimuli. 42

4.1 Predicted means and confidence intervals for proportion of time engaged in the task from models covarying for the age. 51

4.2 Results of participants orienting to name. 52

4.3 Overview of Autism & Beyond App. 62

4.4 Example of automatic coding and results on relationship between age and emotion. 63

4.5 Overview of exhibit space, displayed artwork, and statements for Promotion and Prevention. 73
4.6 Flow of study. ................................................................. 73
4.7 Participants expressed significantly more surprise (and correspond-
ingly, less neutral expression) in Prevention vs. Promotion. ........ 74
4.8 Item recall success and spatial memory accuracy as a function of ex-
pressed surprise and motivational framing. .......................... 74
5.1 Examples of face synthesis methods. ................................. 80
5.2 Organization of collections for base and synthesized face images. . 82
5.3 Visual results for optimal synthesis selections for training and testing. 86
5.4 OFD gallery and probe base images from from a given subject for
testing. ................................. 91
5.5 Rank-1 classification performance as the number of training subjects
increases. ................................. 92
5.6 RoC curves and performance for calibration experiments. ........... 95
5.7 Examples of images and optimal synthesis types from CASIA NIR-VIS
2.0 dataset. ................................. 97
First and foremost, I would like to thank my advisor, Professor Guillermo Sapiro. I owe my deepest gratitude to him for not only being an outstanding mentor, but also a great and compassionate friend. I have learned so much from him and want to thank him for his support, advice, and patience over the years. His constant encouragement, wisdom, and flexibility has allowed me to learn from and pursue a wide variety of topics. It has been an honor working alongside him and he will remain an inspiration for me.

I am also grateful for my graduate committee members, Professor Robert Calderbank, Professor Carlo Tomasi, Professor Larry Carin, and Professor Vahid Tarokh, for their time and effort invested with me, and their helpful feedback. And to Professor Cynthia Rudin, who also provided helpful feedback on my proposal.

Throughout my studies, I have been fortunate enough to collaborate with a wide range of researchers across a multitude of fields. I would like to thank Qiang Qiu, Geraldine Dawson, Kimberly Carpenter, Mariano Tepper, Helen Egger, Kathleen Campbell, Kimberly Chiew, Allison Adcock, and Ricky Bloomfield, for so many great, insightful, and fulfilling discussions.

I would like to give a very special thanks to Qiang Qiu. He has been an amazing friend and great colleague to work with. I am truly grateful for his knowledge and kindness, and cannot emphasize enough how thankful I am of him.

My work has greatly benefited from the amazing people around me, including,
Zhuoqing Chang, Anish Simhal, Steven Espinosa, Natalia Martinez, Martin Bertran, Christopher Tralie, Martin Di Martino, Pablo Sprechmann, Alexey Castrodad, Jinyoung Kim, and Zhongwei Tang. Thanks to them for all of their support, interactions, friendships, and great lunch discussions.

I would also like to thank Kathy Peterson and Amy Kostrewa for their kindness, patience, and help. And to Professor William Engeland, who’s encouragement and kindness helped guide me to pursue an advanced education.

This work would not be possible if it was not for the everlasting support of my family, JoEll, Zack, Dayna, Connor, and Viggo.

It is impossible to remember everyone to thank, and I apologize to those I have inadvertently left out.
Introduction

In today’s world, the information contained in a human face is a valuable commodity. Not only do people care about their own faces, but they hold interest in other’s. With technological advancements, we are getting closer to being able to analyze human faces in ways that were once only possible in sci-fi movies. To name a few: social media organizations are collecting uploaded photos that users have annotated to train systems to automatically recognize people in unseen images (Schroff et al. (2015); Sun et al. (2015)); researchers have shown that aging-rate and health information can be depicted from an image of a person’s face (Belsky et al. (2015)); and psychologists and psychiatrists are able to analyze children’s faces for affect, providing insight to early warning signs of potential developmental disorders (Dawson et al. (2004); Hashemi et al. (2014)). In this dissertation, we bridge a gap between technical contributions and real-world applications of observational behavior analysis. We focus on presenting, validating, and applying novel algorithms to automatically and objectively analyze behaviors. We first present a robust and multi-domain method for pose-invariant facial expression recognition. We then extend this method and present, and validate, other computer vision methods to automatically code behav-
iors (e.g., engagement, name-call responses, and head movement) of toddlers while they watch movie stimuli on mobile devices. We demonstrate real-world impacts of these validated methods by showing relevant findings in screening for autism and cognitive neuroscience using these validated methods. Lastly, we present a framework to calibrate face models to specific deployment domains using data synthesis and augmentation techniques.

1.1 Cross-modality and pose-invariant facial expression recognition

In the first contribution, we develop a robust framework to learn a cross-modality and pose-invariant dictionary representation for facial expression. Pose-invariant facial expression recognition (FER) is the task of classifying facial expressions across various head poses, usually yaw pose changes. For FER, 3D facial features (morphological and shape) are shown to have more discriminative power than 2D facial features (geometric and texture); however generally 3D facial feature information is only available during training and not at testing time. We present a framework that learns using both modalities simultaneously at a multitude of head poses, thus leveraging 3D facial feature information at testing time even when the testing samples are 2D images. We show experimental results and comparisons across publicly available FER datasets.

1.2 Computer vision analysis for quantification of autism risk behaviors

In the second contribution, we develop and deploy a novel mobile application with movie stimuli expertly curated to elicit affective behaviors and social responses in toddlers. Currently, identification of children with neurodevelopmental disorders in early childhood requires low specificity screening with questionnaires followed by time-consuming, in-person observational analysis by highly-trained experts. To this
end, we developed this mobile application and computer vision algorithms to automatically analyze key behavioral markers including engagement, name-call responses, and emotional responses. Results of the presented methodology and automated analysis are shown on a population of children with and without autism and validated across expert human coders.

1.3 Findings and extensions using computer vision analysis

In the third contribution, we present findings from research studies where the developed methods were deployed. We first present two studies focused around quantification of autism spectrum disorder risk behaviors of toddlers and young children in clinic and in at-home environments. Then, a study focused on a novel investigation into motivated exploration and memory for a real-life, naturalistic environment.

1.4 Framework for intelligent synthesis driven model calibration

In the final contribution, we explore a future direction of the developed methods: how to efficiently use face synthesis and augmentation to calibrate pre-trained face models to specific deployment domains. State-of-the-art face analysis methods are prone to suffer severe performance issues when deployed in many real-world scenarios. These performance issues stem from shifts between the training and deployment domains such as image resolution, lighting conditions, occlusions, ethnicities of subjects, among others. Pre-trained models can be adapted to a given deployment domain; furthermore, when only a limited amount of deployment data is available, face synthesis methods enrich the limited data while maintaining the deployment domain. We propose multiple face synthesis methods and information-driven approaches to exploit and optimally select face synthesis methods and samples to calibrate state-of-the-art face analysis models to different deployment scenarios. We show that our approaches lead to more efficient training and improved testing performances.
1.5 Organization of the dissertation

In Chapter 2, we discuss contributions to cross-modality and pose-invariant facial expression recognition. In Chapter 3, we introduce a closed-loop, self-contained, mobile application to analyze early behavioral risk markers for autism. In Chapter 4, we present findings using the developed methods, including group differences for behavioral risk markers for autism and interactions between motivational framing context, facial affect, and memory outcome. Finally, in Chapter 5, a framework for calibrating pre-trained models for face analysis tasks is presented.
2

Cross-modality and pose-invariant facial expression recognition

2.1 Chapter summary

In this Chapter, we present a dictionary learning based framework for robust, cross-modality, and pose-invariant facial expression recognition. The proposed framework first learns a dictionary that i) contains both 3D shape and morphological information as well as 2D texture and geometric information, ii) enforces coherence across both 2D and 3D modalities and different poses, and iii) is robust in the sense that a learned dictionary can be applied across multiple facial expression datasets. We demonstrate that enforcing domain specific block structures on the dictionary, given a test expression sample, we can transform such sample across different domains for tasks such as pose alignment. We validate our approach on the task of pose-invariant facial expression recognition on the standard BU3D-FE and MultiPie datasets, achieving state of the art performance.
2.2 Introduction and related work

The analysis of facial expression is studied in computer vision, psychology, psychiatry, and marketing, all of which require a facial expression recognition (FER) system to be robust to changes in pose. In particular for the psychology and psychiatry fields, risk signs of anxiety and autism can be depicted from facial expressions as the participant is looking at various stimuli (Nichols et al. (2014); Ozonoff et al. (2010)). Robustness to pose is especially important since the experts need to analyze participants in their natural states, in other words being observed in an unconstrained manner (see Rehg et al. (2013) and Hashemi et al. (2014) for examples). Many state of the art facial expression approaches focus on frontal or nearly frontal images of the face (Shan et al. (2009); Zeng et al. (2009)). Changes in head pose or facial expression cause nonlinear transformations of the face in a 2D image, making it a non-trivial task to classify expressions under varying poses (Zhu and Ji (2006)). Even with recent FER advancements, manually coding of facial expression is still performed in the psychiatry and psychology fields due in part to this challenge.

Approaches to handle facial expression across multiple poses fall within two main categories. The first category corresponds to approaches based on learning expression models on a discrete set of poses (Moore and Bowden (2011); Tang et al. (2010)). For example, Moore and Bowden (2011) employ a 2 stage approach where they first train a classifier to distinguish pose, and then train pose-dependent classifiers across expressions. The second category involves approaches that learn the mappings of the expressions as a function of pose (Rudovic et al. (2013); Kumano et al. (2009); Guney et al. (2013)). Notably, Rudovic et al. (2013) presents an accurate geometric based approach to first learn the transformation of facial points at any given pose to a frontal pose, then FER is performed on facial points from the projected frontal pose, thus requiring only one posed classifier. The work Guney et al. (2013) adopts
a Partial Least Squares approach, that has been explored in facial recognition, to model the relations between pairs of images of the same person at different poses and expressions.

In addition to FER in 2D images, much attention has been focused on using 3D face scans (Zhao et al. (2013); Sandbach et al. (2012)). Specifically, textured 3D face scans not only contain 3D features (e.g., morphological and shape), but also 2D features (e.g., geometric and texture). Zhao et al. (2013) have shown that when dealing with 2D and 3D features independently on a frontal face, the ordering of discriminative power for FER is morphological, shape, and texture; and combining all three feature modalities together achieves the strongest discriminative power. Although textured 3D face scans provide the most discriminative features, technology has not yet allowed for practical acquisition in unconstrained environments, such as capturing child facial behaviors in a doctor’s office.

Dictionary based approaches have been extensively used for classification and regression in the areas of facial recognition and expression (Taheri et al. (2014); Wright et al. (2009)). Furthermore, one can apply sparse based methods by incorporating regularized penalty functions to determine sparse coefficients in a more greedy fashion (Wright et al. (2009); Hastie et al. (2001)). By encoding structure along atoms in the dictionary, such as annotating or grouping atoms in the dictionary with class labels, the sparse coefficients can provide knowledge to the class that the unseen face belongs to. Recent work has also focused on encoding structure within the atoms themselves, namely domain adaptive dictionary learning (Qiu et al. (2012)). A powerful example focuses on encoding atoms so they contain blocks of features across different domains (Qiu et al. (2012)).

We develop a framework based on learning and applying a robust, cross-modality, and pose-invariant dictionary to the recognition of facial expressions. The presented framework first learns a dictionary that i) contains both 3D shape and morphological
Figure 2.1: Overview of multi-modality and pose-invariant dictionary construction. (a) Samples of 3D textured face scans from the BU3D-FE dataset, where each sample $B\{n,m\}$ can be decomposed into a 3D and 2D component, $b_t\{n,m\}$ and $b_0\{n,m\}$ respectively. The used 19 facial landmarks are highlighted with green markers. (b) The dictionary is composed of blocks containing different modalities and poses. The red sections represent dictionary block $D(b_0)$ containing 3D features, while the blue sections represent dictionary blocks $D(b_t)$ containing 2D features from the synthesized head poses $t$.
information as well as 2D texture and geometric information, ii) enforces coherence across both 2D and 3D modalities and different poses, and iii) is robust in the sense that a learned dictionary can be applied to multiple facial expression datasets. With our novel dictionary based approach, we achieve powerful results in the task of pose-invariant FER.

In this chapter we make the following contributions:

- We learn a robust, cross-modality and pose-invariant dictionary for facial expressions.
- We describe how the learned dictionary leverages 3D information at testing time even when the samples at testing are 2D images.
- We achieve state-of-the-art performance on pose-invariant FER and demonstrate how it can be applied across different datasets.

The rest of the chapter is organized as follows: in Section 2 we describe our approach for constructing and applying the proposed dictionary to pose-invariant FER. In Section 3 we validate our approach using two publicly available datasets: the BU-3D Facial Expression (BU3D-FE) (Yin et al. (2006)) and the CMU Pose, Illumination, and Expression (MultiPie; Gross et al. (2010)). Section 4 concludes the chapter.

2.3 Proposed framework

We now describe our approach, which can be separated into two main components: constructing the cross-modality and pose-invariant dictionary, and applying it to the task of pose-invariant FER. Figures 2.1 and 2.2(a) show the outline of our approach for dictionary construction and cross domain representation.
Figure 2.2: (a) Overview of the proposed cross-modality and pose-invariant representation and examples from the MultiPie dataset. Given the dictionary $\tilde{D}$, we propose that the sparse coefficient vectors between the same subjects performing the same expressions at different poses or modalities will be nearly identical. The dotted colored boxes around the faces (orange and green) represent observation of the same subjects from a given expression at different poses. These observations can be represented by a linear combination of the same sparse coefficients being applied to a given sub-dictionary of $\tilde{D}$, that is represented with the respective dotted color boxes. (b) Example of the sparse coefficient vectors extracted from a subject performing a Disgust expression at 7 poses. Each of the 7 columns in the image correspond to a sparse coefficient vector $x$ extracted at a given pose, and the rows represent weights corresponding to atoms in $\tilde{D}$. Images of the input subject are shown below each pose. Notice the horizontal line structure (depicted by the red arrows) throughout the sparse coefficient vectors at different poses, reinforcing the notation that the sparse coefficients extracted for different poses are approximately consistent thus pose invariant.
2.3.1 Pose-invariant dictionary

Given a dataset containing 3D textured face scans of \( N \) subjects across \( M_e \) expressions, we define each sample as \( B\{n,m\} \) where \( n = 1, \ldots, N \), and \( m = 1, \ldots, M_e \) represent the subject and expressions labels respectively. From a single 3D textured face scan, 2D images with varying head poses \( t \) can be synthesized. In this sense, we can decompose a sample as

\[
B\{n,m\} = b_0\{n,m\} + b_t\{n,m\} \quad \text{for} \quad t = 1, \ldots, T
\]

where \( b_0\{n,m\} \) represents the 3D specific information and \( b_t\{n,m\} \) for \( t = 1, \ldots, T \) represents the 2D specific information across \( T \) different poses. Note that 3D specific information does not change with varying head poses, and \( t = 1 \) represents a 2D frontal face image. By aggregating across all samples, we define the dictionary block \( D(b_0) \in \mathbb{R}^{d_m \times (N \times M_e)} \) of extracted 3D features as

\[
D(b_0) = [b_0\{1,1\}, \ldots, b_0\{n,m\}, \ldots, b_0\{N,M_e\}]
\]

where \( b_0\{n,m\} \in \mathbb{R}^{d_m} \) represents the column array of computed frontal 3D features from the sample of subject \( n \) performing expression \( m \). Similarly, for all samples with simulated head pose \( t \), we define the block \( D(b_t) \in \mathbb{R}^{d_n \times (N \times M_e)} \) of extracted 2D features as

\[
D(b_t) = [b_t\{1,1\}, \ldots, b_t\{n,m\}, \ldots, b_t\{N,M_e\}]
\]

where \( b_t\{n,m\} \in \mathbb{R}^{d_n} \), represents the column array of computed 2D features from the sample at pose \( t \).

The cross-modal and pose-invariant dictionary, \( \mathcal{D} \), is organized by stacking the dictionary blocks (see Figure 2.1)

\[
\mathcal{D} = [D(b_0); D(b_1); D(b_2); \cdots ; D(b_T)],
\]

with the stacking operator \([D(b_0); D(b_t)] = \begin{bmatrix} D(b_0) \\ D(b_t) \end{bmatrix}\). \( \mathcal{D} \in \mathbb{R}^{(d_m + T \times d_n) \times (N \times M_e)} \) is composed of a total of \( T + 1 \) blocks, specifically one block containing the 3D features,
D(b_0), and T blocks containing the 2D features from each of the T simulated head poses, D(b_t). This block structure within the dictionary D imposes coherence across the different domains.

In addition, we learn a more compact dictionary by applying a dictionary learning method, such as K-SVD (Cai et al. (2014)), creating a new dictionary \( \tilde{D} \in \mathbb{R}^{(d_m+T \times d_n) \times d_d} \) where \( d_d \leq N \). Note that since the block structure still remains, the coherence between the domains is preserved: D is transferred to
\[
\tilde{D} = [\tilde{D}(b_0); \tilde{D}(b_1); \tilde{D}(b_2); \cdots; \tilde{D}(b_T)],
\]
where now \( \tilde{D}(b_0) \in \mathbb{R}^{d_m \times d_d} \) and \( \tilde{D}(b_t) \in \mathbb{R}^{d_n \times d_d} \) (see Figure 2.1).

2.3.2 Cross modality and domain representation

The learned dictionary, \( \tilde{D} \), contains a dense amount of expression information jointly learned across multiple domains (3D and different poses in 2D). Let unseen samples containing expression class labels and only 2D images at any pose \( t \) be defined as \( q_t \{n,m\} \). The goal is to represent \( q_t \) as a sparse linear combination of the frontal 3D and frontal 2D features in \( \tilde{D} \), namely \( [\tilde{D}(b_0); \tilde{D}(b_1)] \), since they are known to have large discrimination power for FER (for ease of reading, \( q_t \) is short-hand for \( q_t \{n,m\} \)). Thus we wish to solve:

\[
\begin{align*}
\{\tilde{q}_t, x\} &= \arg \min_{x,\tilde{q}_t} \|\tilde{q}_t - [\tilde{D}(b_0); \tilde{D}(b_1)] x\|^2_2 \\
\text{s.t.} \|x\|_0 &\leq \lambda,
\end{align*}
\]

(2.1)

where \( x \in \mathbb{R}^{d_d} \) is the sparse coefficient vector, \( \tilde{q}_t \in \mathbb{R}^{d_n+d_m} \) is the transformed version of sample \( q_t \) onto the domains represented by \( [\tilde{D}(b_0); \tilde{D}(b_1)] \), \( \|x\|_0 \) counts the number of non-zeros values in \( x \), and \( \lambda \) is the imposed sparsity constant. Equation (2.1) is not directly solvable since the 3D information and frontal 2D information, \( \tilde{q}_t \), and the sparse coefficient vector, \( x \), are unknown. Instead we propose to represent the unknown 3D and frontal 2D information via our domain adaptive dictionary. We
propose that the computed sparse coefficient vector in the known domain provided by \(q_t\) can be directly applied to dictionary blocks in unseen domains to estimate \(\tilde{q}_t\).

Since \(q_t\) provides information in the same domain the dictionary block \(\tilde{D}(b_t)\), the sparse coefficient vector can be determined from:

\[
x = \arg\min_x \|q_t - \tilde{D}(b_t) x\|_2^2, \quad s.t. \|x\|_0 \leq \lambda.
\]

If \(t\) is unknown, it can be estimated from a variety of head pose approaches (Chutorian and Trivedi (2009)) or by determining which domain block in \(\tilde{D}\) gives the lowest reconstruction error. Due to the coherence across domains within the stacks of the dictionary \(\tilde{D}\), we assume the sparse coefficient vector, \(x\), should not differ greatly between extracted data of the same subject but in different domains (see Figure 2.2(a)). In other words,

\[
x = \arg\min_x \|q_t - \tilde{D}(b_{t'}) x\|_2^2, \quad s.t. \|x\|_0 \leq \lambda
\]
\[
\approx \arg\min_x \|q_{t'\neq t} - \tilde{D}(b_{t'\neq t}) x\|_2^2, \quad s.t. \|x\|_0 \leq \lambda
\]
\[
\approx \arg\min_x \|q_0 - \tilde{D}(b_0) x\|_2^2, \quad s.t. \|x\|_0 \leq \lambda.
\]

This assumption is explored and further validated in Section 2.4.2 and Figure 2.2(b).

Equation (2.2) states that \(x\) can be determined from any domain that lies in both \(q_t\) and \(\tilde{D}\). Once \(x\) is determined, \(\tilde{q}_t\) can be computed from \(2.1\).

### 2.4 Experimental validation

#### 2.4.1 Datasets and setup

To evaluate our proposed method, we used two publicly available face datasets: the BU3D-FE and the MultiPie datasets. The BU3D-FE dataset consists of textured 3D face scans of 100 subjects performing 6 different expressions: Anger (AN), Disgust (DI), Fear (FE), Happy (HA), Sad (SA), Surprised (SU) at 4 different levels, and a
Neutral (NE) expression (see Figure 2.1 for examples). For this demonstration, we only considered the data from the maximum level which corresponds to the apex of the expression. From the MultiPie dataset, we selected 2D images from 160 subjects performing 4 different expressions: DI, HA, SU, and NE at 7 different yaw angles (0, -45, -30, -15, 15, 30, 45) (see Figure 2.2(a) for examples). The MultiPie dataset also contains each expression and pose at different illuminations; however, we only considered the data from the frontal illumination.

To compute features from the datasets, 49 facial landmarks were automatically extracted with the IntraFace software (De la Torre et al. (2015)). Faces were aligned and normalized to a mean face across the BU3D-FE dataset using the inner-eye landmarks and the spine of nose. For selection of 2D and 3D features, we followed the state of the art approach in Zhao et al. (2013), where four modalities of features consisting of morphological, shape, texture, and geometric features are computed around 19 of the facial landmark points (see Figure 2.1). 3D morphological features consist of 157 Euclidean distance pairs between the 19 landmarks on the range data of the faces. 3D shape features consist of multi-scale local binary pattern (LBP) patches around each of the 19 landmarks on the image of the range data. Specifically, we compute LBP with radii ranging from 1 to 5, where the total features extracted across all the patches at a given LBP scale is 4275. 2D texture features are computed in the same manner as the 3D shape features except extracted on the 2D textured images. 2D geometric features consist of the same distance pairs as the 3D morphological features, but the range value of each landmark is not considered. Principal component analysis (PCA) is performed on each modality independently, preserving at least 95% of the variation, thus reducing the dimensions of the morphological, shape, texture, and geometric features to 100, 1000, 1000, 100 respectively. Thus in the following experiments $d_n = d_m = 1100$. For all experiments shown, 2D images containing 7 poses with yaw angles (0, -45, -30, -15, 15, 30, 45) were considered.
Table 2.1: Comparisons of recognition rates (%) for varying expressions across different methods on BU3D-FE dataset, including a 3D specific framework (Zhao et al. (2013)), a pose-invariant framework (Rudovic et al. (2013),) and our proposed approach when Neutral is and is not included. Note that Zhao et al. (2013) only considers a frontal pose and use 3D data for testing, while we adopt a more general and challenging testing setup.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Expression</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AN</td>
<td>DI</td>
</tr>
<tr>
<td>3D-Zhao et al. (2013)</td>
<td>83</td>
<td>87</td>
</tr>
<tr>
<td>Proposed</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>2D-Rudovic et al. (2013)</td>
<td>68</td>
<td>75</td>
</tr>
<tr>
<td>Proposed</td>
<td>82</td>
<td>85</td>
</tr>
<tr>
<td>Proposed w/NE</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The sparse coefficient vectors for the projected discriminant frontal $\tilde{q}_t$ representations were determined through Orthogonal Matching Pursuit (OMP) with sparsity constant $\lambda = \frac{1}{7} d_d$, since the dictionary is composed of an even representation of samples across 7 expressions. For each experiment, a single facial expression classifier was trained by applying to the extracted $\tilde{q}_t$ representations a multi-class Support Vector Machine (SVM) (Chang and Lin (2011)) with a radial basis function kernel. Experiments in Section 2.4.2 perform a five-fold cross validation procedure to construct and test on the pose-invariant dictionary. Out of the samples chosen for constructing the dictionary, a ten-fold cross validation procedure was performed to determine the optimal SVM parameters.

2.4.2 Validation on BU3D-FE

We now present experiments performed on the BU3D-FE dataset. Since the 3D modalities and the 7 poses are considered for the dictionary, it contains 8 dictionary blocks (see Figure 2.1). Furthermore, K-SVD was applied to create a compact dictionary $\tilde{\mathcal{D}} \in \mathbb{R}^{8800 \times 400}$. For testing, 2D images of expressions performed at the 7 pose angles are used. Figure 2.2(b) provides an example of the sparse coefficients vectors.
extracted from a given subject performing a specific expression at multiple poses. In this figure one can observe many sparse coefficients that are present throughout all of the poses (red arrows), thus illustrating that the learned dictionary is invariant to observed poses. Furthermore since we assume the sparse coefficient vector is approximately the same given any modality from a sample, then we can project a given sample to modalities that may not have been observed (e.g., projecting a posed image to one containing 3D features).

Our approach achieved high results for pose-invariant FER, achieving 82% and 85% recognition rates when Neutral is and is not considered. In Table 2.1 we compare our results to those of two recently published, state of the art methods, namely a pose-invariant method involving only 2D modalities (Rudovic et al. (2013)) and a 3D specific method that only considers frontal face scans (Zhao et al. (2013)). It should be noted the testing setup differed between cited methods. Rudovic et al. (2013) provide results using manually annotated facial landmarks, and test on a wide variety of poses unseen to the training data including pitch poses. Zhao et al. (2013) only consider 3D face scans, frontal pose, and do not classify the Neutral expression. With this said, our proposed approach therefore achieves results for FER that are on par with current state of the art approaches on the BU3D-FE dataset in a more general and challenging setting. When not including the (challenging) Neutral expression, we achieve the same recognition rate as Zhao et al. (2013) even though they only use the frontal pose and 3D data for testing.

2.4.3 Pose-invariant FER from frontal training data

We now present an experiment that utilizes both the BU3D-FE and MultiPie datasets, in order to validate the robustness of our approach. We propose to first learn a cross-modal and pose-invariant dictionary with the textured 3D data provided by the BU3D-FE dataset. Then using the 2D images from the MultiPie dataset, we
Table 2.2: Comparisons of recognition rates for all expressions across the 7 poses on the MultiPie dataset. We consider two proposed versions: one with head pose explicitly provided and one without. Both proposed methods perform consistently well across drastic pose changes and significantly outperform the baseline at severe pose angles.

<table>
<thead>
<tr>
<th>Pose (deg)</th>
<th>-45</th>
<th>-30</th>
<th>-15</th>
<th>0</th>
<th>15</th>
<th>30</th>
<th>45</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>67</td>
<td>76</td>
<td>90</td>
<td>91</td>
<td>91</td>
<td>75</td>
<td>64</td>
</tr>
<tr>
<td>Proposed w/out head pose</td>
<td>80</td>
<td>82</td>
<td>89</td>
<td><strong>92</strong></td>
<td><strong>90</strong></td>
<td>84</td>
<td>81</td>
</tr>
<tr>
<td>Proposed w/head pose</td>
<td><strong>86</strong></td>
<td><strong>87</strong></td>
<td><strong>90</strong></td>
<td><strong>92</strong></td>
<td><strong>90</strong></td>
<td><strong>88</strong></td>
<td><strong>85</strong></td>
</tr>
</tbody>
</table>

wish to train and test a FER classifier. Furthermore, we wish to demonstrate the power of our proposed dictionary formulation by learning pose-invariant FER classification models using only frontal faces from the MultiPie dataset as training samples and testing on posed 2D images from the MultiPie dataset. This is the first instance where both of these datasets are utilized simultaneously and in this fashion. Although the expressions presented in the MultiPie dataset are only a subset of those presented in the BU3D-FE dataset, we trained a pose-invariant dictionary based on the entire BU3D-FE at the 7 dynamic pose angles to demonstrate our approach’s general usability. Similar to the experiments in Section 2.4.2, we applied K-SVD to get a final pose-invariant dictionary $\tilde{D} \in \mathbb{R}^{8800 \times 400}$. Five-fold cross validation is carried out on the MultiPie dataset, where at each fold 80% of the MultiPie subjects are used to train a classifier and the other 20% of the subjects at 7 different poses are used for testing. The dictionary learned from the BU3D-FE dataset did not change throughout any of the folds.

Table 2.2 shows the total recognition rates for all expressions across each of the 7 poses for two versions of our proposed method and a baseline method. The proposed methods in the previous setup assumed head pose information was explicitly provided at testing, in this experiment we explored the contributions of providing head pose at testing versus not for pose-invariant FER, and compared results to a baseline method.
that consisted of training a multi-class SVM for each of the expressions performed on a frontal pose using the same set of 2D features as in Section 2.4.1. For a testing with images that do not have head pose explicitly defined, we determined which domain block the coefficient vector should be estimated from by iteratively computing the reconstruction errors across all domain blocks in $\tilde{D}$ and choosing the one which gave the lowest error. All methods perform very well for nearly frontal faces when the pose is between $-15$ and $15$ degrees; however outside this range, as severe pose changes occur, our methods greatly outperform the baseline method. For drastic pose angles, providing head pose at testing achieves highest performance with recognition rates similar to those of nearly frontal faces.

Figure 2.3 shows the confusion matrix result when only frontal data from the MultiPie dataset and the learned dictionary from the BU3D-FE dataset are used to train a classifier, testing is performed on the 7 dynamic poses and head pose is provided at testing. Our approach achieved a high FER rate of $87\%$ across the varying poses, further validating the value of the proposed dictionary to cross-modality and
pose-invariant representation of facial expressions.

2.5 Conclusion

We have presented a framework for constructing and learning a cross-modality and pose-invariant dictionary for the task of facial expression recognition. Using the BU3D-FE dataset, we have shown we get results on par with current (frontal) state of the art approaches for 3D and pose-invariant expression recognition. Furthermore, we have validated the robustness of our approach by achieving high performance when two different datasets are combined. The generic nature of our approach allows for many extensions including the use of different features and modalities.
3

Computer vision analysis for quantification of autism risk behaviors

3.1 Chapter summary

In this Chapter, we present a self-contained, closed-loop, mobile application with movie stimuli designed to engage the child’s attention and elicit specific behavioral and social responses, which are recorded with a mobile device camera and analyzed via computer vision algorithms. In addition to presenting this paradigm, we validate the system to measure engagement, name-call responses, and emotional responses of toddlers with and without ASD who were presented with the application. We also show examples of how the proposed framework can further risk marker research with fine-grained quantification of behaviors.

3.2 Introduction

Observational behavior analysis has played, and still plays, a key role for gaining insight into mechanisms and risk markers of impairing neurodevelopmental disorders such as autism spectrum disorder (ASD). The gold standard observational tool for
ASD diagnosis, the Autism Diagnostic Observational Schedule (ADOS, Lord et al. (2000)), requires several observational coding components, as does an assessment of early risk markers of ASD, the Autism Observational Scale for Infants (AOSI, Bryson et al. (2008)). Retrospective behavioral analysis of home videos helped discover early risk markers involving diminished social engagement and joint attention in children who were later diagnosed with ASD (Adrien et al. (1991, 1992); Werner and Dawson (2005); Mars et al. (1998); Osterling and Dawson (1994)). Research studies have documented several behavioral risk markers by showing that toddlers with ASD exhibit deficits in orienting to name calls, range of affect, social smiles, and other aspects of visual attention, (e.g., Elsabbagh et al. (2013); Ozonoff et al. (2010); Zwaigenbaum et al. (2005); Martin et al. (2018)). These advancements in understanding early behavioral development aid in the development of tools for screening, diagnosing, and monitoring ASD.

Studies and evaluations involving behavioral analysis tend to rely heavily on medical practitioners and specialists who have undergone intensive training to be able to reliably administer the eliciting tasks and then code and interpret the observed behaviors. In behavioral studies, practitioners tend to code behaviors based on clinical judgment and thus tend to code behaviors at a lower granularity than is possible in computer analysis. In retrospective analysis, specialists can potentially go frame-by-frame to hand code behaviors; this is not only burdensome, but is not easily scalable for big data studies aimed at discovering or refining behavioral risk markers or for longitudinal tracking. As technology has advanced, new tools have emerged to assist in automatic and semi-automatic behavioral coding in infants and toddlers. Eye tracking is a great example, where technological advancements have made major impacts in the understanding of ASD behavioral development. Researchers are able to automatically code behavior from gaze and analyze it at fine-grained scales, leading to novel understandings of development (Constantino et al. (2017); Falck-Ytter
et al. (2013)). However, standard eye tracking systems are still very constrained, specialized, and expensive, making the availability and reach of these systems limited. Research using automatic behavioral coding in less constrained settings, from schools to homes, have also been promising, where for example researchers explore tools and a dataset to develop and evaluate social and communicative behaviors relevant to child-adult interaction sessions (Rehg et al. (2013)). In another recent work, researchers automatically encoded motor movements during mother-infant interactions to explore quality of interaction (Egmose et al. (2017)). There also has been progress in augmenting the codings for visual attention tasks in infants with the use of just a single consumer grade camera (Hashemi et al. (2014)), and automating coding of head movement dynamics while watching movies of nonsocial and social stimuli (Martin et al. (2018)).

In this work we concentrate on the use of ubiquitous devices, like smart phones and tablets, developing a solution that doesn’t need any additional hardware at all. Focusing on unconstrained and low-cost setup requirements is especially important since many middle and low resource communities lack access to specialists in ASD, thus lack access to any sort of evaluation. Additionally, unconstrained setups allow for behavioral monitoring in more naturalistic environments, such as at home. There is a desire for universal and well-validated behavioral tools to further research on early risk marker detection in ASD. The tools have to be universal in the sense that they are consistent and accessible across different user groups; and have to be validated in the sense that they have to agree with trained specialists, potentially helping to discover new biomarkers as further discussed in this paper. It is worth noting that ASD diagnosis involves much more than the detection of risk markers, but furthering the accessibility and development of screening tools for identifying toddlers that might be at risk and informing caregivers, is of great value. Although risk behaviors can be observed as early as at 12 months of age leading to diagnosis
at 18 months, the average age of diagnosis is the United States is around \( \sim 4-5 \) years old (Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators, Centers for Disease Control and Prevention (CDC) (2014)). Not only can aggressive early intervention provide long-term improvements for the child, but also starting intervention before the full syndrome is present can have an even greater effect on outcomes (Pickles et al. (2016); Jones et al. (2017); Webb et al. (2014)).

Towards these challenges and opportunities, our interdisciplinary team developed a mobile application with movie stimuli designed to elicit and quantify specific ASD-related behavioral responses in toddlers. In a long-term effort, this project aims to develop low-cost, automatic, and quantitative tools that can be used by researchers and general practitioners in general settings (such as clinics or schools, and potentially by caregivers at home) to identify toddlers at risk for ASD or other developmental disorders.

We have developed a novel application of displaying movie stimuli on a mobile device which were expertly curated to capture the toddler’s attention and elicit relevant behaviors to early risk markers of ASD, including orienting to name call, social referencing, smiling, pointing, and social smiling (Ozonoff et al. (2010); Zwaigenbaum et al. (2005, 2015); Dawson et al. (2004)). Using the front facing camera of the mobile device, we capture the toddler and automatically code these event-related behaviors with computer vision analysis tools. The current study presented in this paper validates our automatic codings of engagement, name-call responses, and emotion with hand codings from specialists on a diverse population including toddlers with ASD and without ASD (non-ASD). We also present examples of behavioral analysis extracted from the validated automatic methods.

It is important to stress that while the framework here introduced can be used in any environment, from clinics to homes, it is not a passive monitoring system but
a user friendly and active one. In passive systems, e.g., those used to monitor heart conditions via wrist watches, the signal to noise ratio is very poor, in particular when aiming at capturing highly accurate and risk informative behaviors within a short time span. With a closed loop system like the one here proposed, where carefully designed short and entertaining movie stimuli elicit behaviors (as in ADOS or AOSI standard approaches, but without the human in the loop), we obtain a much more valuable and interpretable signal.

3.3 Data collection

3.3.1 Setup and participants

The study was carried out in a pediatric care clinic with the approval of the Duke Health Institutional Review Board. The caregivers of toddlers visiting the clinic for an 18 or 24 month well-child visit, where all toddlers in the clinic are screened for ASD with The Modified Checklist for Autism in Toddlers – Revised with Follow-up (M-CHAT-R/F) Robins et al. (2015), were invited to participate in the study. Children with known hearing or vision impairments and caregivers who could not complete consenting in English were excluded. During the well-child visit, caregivers were asked to hold the toddler on his/her lap while an iPad (4th generation) was placed on a stand at the toddler’s eye-level and set about 1 meter away, see Figure 3.1. To minimize distractions, all other family members, persons, and the practitioner were asked to stand behind the caregivers and toddler. Caregivers were told they could interact with their toddler for the first 45 seconds of the session while a ‘mirror’ was presented on the screen (child can see him/herself on the screen, which is showing the tablet’s camera input). After this time, the caregivers were asked to remain quiet and not direct their child’s behavior or attention once the movie stimuli began. For the rest of the session, the iPad displayed different carefully designed movie stimuli on the screen (one stimulus at a time, with nothing else shown on the screen), while
Table 3.1: Demographics table reported in means (standard deviations) or counts (percentage)

<table>
<thead>
<tr>
<th></th>
<th>non-ASD</th>
<th>ASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total participants</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>Males</td>
<td>16 (89%)</td>
<td>13 (87%)</td>
</tr>
<tr>
<td>Age in months</td>
<td>26.4 (std 4.6)</td>
<td>25.5 (std 3.8)</td>
</tr>
<tr>
<td>ADOS-T Total</td>
<td>-</td>
<td>17.7 (std 5.3)</td>
</tr>
</tbody>
</table>

simultaneously recording the child’s face via the front facing camera at 1280x720 resolution and 30 frames per second. If a child screened positive on the M-CHAT-R/F, or a caregiver or clinician expressed concerns about ASD during the visit, then the child received gold standard diagnostic testing with the Autism Diagnostic Observational Schedule – Toddler Module with a child psychologist for final diagnosis (Luyster et al. (2010)). As a validation set for the developed computational methods, we consider a total of 33 participants (18 non-ASD and 15 ASD) ages 16 - 31 months enrolled in our study. Full demographics are shown in Table 3.1.

### 3.3.2 Movie stimuli

A series of stimuli were presented on the iPad screen, consisting of short developmentally appropriate movies designed to engage the child’s attention and elicit specific behavioral and social responses. The movies incorporated a ‘name-call’ prompt where during the movie the practitioner, who was standing behind the child and caregiver, would call the child’s name loudly while the movie is still being displayed on the screen (this is of course easy to incorporate in an automatic fashion, e.g., via bluetooth speakers). Screenshots of the stimuli and the recording from the iPad are shown in Figure 3.1. We considered the set of 3 movie stimuli, namely:

**Bubbles.** Bubbles are presented at random and moving throughout the frame. A ‘name-call’ is presented once during the movie. The total duration is 30 seconds.

**Bunny.** A mechanical toy bunny is presented on one side of the screen and hops
Figure 3.1: Screenshot of the recorded video from the front facing tablet’s camera and example of automatic facial landmarking are shown in first row. In this screenshot, the child (1) is sitting on the caregiver’s (2) lap, while the practitioner (3) is standing behind. All six outlined automatically detected landmarks (in black) are used for face pre-processing, while the lowest nose and the two outer eye landmarks are used to track head movement. Screenshots of frames from the movie stimuli being presented are shown in the remaining rows. These are Bubbles, Bunny, and Puppets, respectively.
horizontally towards the other side, which contains a group of toy vegetables. As the bunny reaches the midpoint of its path, an animal puppet is introduced and temporarily disrupts the bunny’s path. A ‘name-call’ is presented once during the movie. The total duration is 66 seconds.

**Puppets.** Two animal puppets interact while building a block tower together and then knock it down. The tower is built three times and knocked down twice. A ‘name-call’ is presented once during the movie. The total duration is 68 seconds.

3.4 Methods

We developed automatic video and audio analysis methods to study behaviors and child responses related to attention and emotion. During the movie stimuli, the text ‘Name’ appeared in the upper left corner to prompt the practitioner to call the child’s name once loudly (see previous section). Although it is known when the text appeared on the screen in relation to the movie, there is a need for automatic detection of the exact time the practitioner said the child’s name. To this end, we also developed an automatic name-call detection based on the recorded audio. To validate the automatic methods, expert human raters coded emotion, name-call prompts, and name-call responses. While the computer vision and machine learning communities have now very advanced ground-truth data, including for human emotions, this is restricted to mostly healthy adults. It is therefore imperative to validate results with the population of interest, since it is well documented that children, and also children with developmental disorders, exhibit very different facial expressions when compared with such standard database populations.
3.4.1 Automatic coding

Name-call prompt detection

Since we know when in the movie stimuli the practitioner is prompted to say the child’s name for a name-call, a 4 second window of the recorded audio data is extracted around each of the known name-call text prompts starting 1 second before the time when the text prompt occurred. Focusing on the power spectrum density (psd) of the extracted audio signal in the 300–1,200 Hz frequency range, we estimate when the adult practitioner said the child’s name by finding the time corresponding to maximum speech energy, see Figure 3.3.

Face pre-processing

A facial landmark detector and tracker was deployed to track 49 facial landmark points De la Torre et al. (2015). While other algorithms could be used without affecting the proposed paradigm, this one was found to be very efficient for the desired tasks. Using a subset of the landmark points, namely the inner and outer eye points and the nose spline (see Figure 3.1), faces were aligned and normalized to a frontal canonical face model by finding the affine transformation between them. In turn, this transformation also represents the head pose estimation. The facial landmark detector requires both eyes to be present in the frame, thus the range of estimated yaw pose $\theta_{\text{yaw}}$ is $\{-45^\circ, +45^\circ\}$ (left-right head orientation). We assume that the toddler’s head orientation is directly correlated to if he/she is watching towards the stimuli. This assumption is supported by the ‘center bias’ property that is well established in gaze estimation literature Manna et al. (1995); Li et al. (2013b). Engagement when the toddler is watching towards the movie stimuli is defined by frames when the toddler exhibits yaw poses with magnitudes less than 20°.
Figure 3.2: Example of a head turn using the automatic method. To differentiate a head turn from a face occlusion, we determine if the child is performing a head turning motion before and after the face is lost or when its exhibiting a yaw pose with large magnitude. The red bars represent the half-second windows used to determine if the child is exhibiting a head turning motion before and after the face is lost (by the camera) or when its exhibiting a yaw pose with large magnitude.

Figure 3.3: Audio is analyzed to determine the exact time point the practitioner said the child’s name during a name-call. The power spectrum density (psd) of the recorded audio signal (3.3(a)) contains audio from the movie stimuli (predominantly music) and instances of vocalizations. Root mean squared (RMS) values of the audio signal (3.3(b)) provide quantification of audio signals at each time point, and are used to detect a name-call prompt. Knowing that practitioner was asked to prompt a name-call at 15 seconds into the stimuli, in this example we are able to focus on speech around the time point (green box) and detect the exact time point when maximum speech occurred.
Head movement and turn detection

We estimate the child’s head movement by tracking the distances and pixel-wise displacements of central facial landmarks. We record the frame-by-frame displacements of landmarks around the nose, namely the two outer eye landmarks and the lowest nose landmark shown in Figure 3.1. The magnitudes of these displacements are heavily dependent on the distance the child is away from the camera. Thus these displacements need to be normalized with respect to the child’s distance from the camera. If depth information were available, this would be a trivial task; however, since it is not, we normalize the displacements with respect to the distance between the child’s eyes, keeping in line with the use of only available and ubiquitous hardware. At any given time point, the displacements from the nose landmark are normalized by a $\pm 1$ second windowed-average Euclidean distance between the eyes.

Since the practitioner and caregiver are located behind the child, the child must transition his/her face from looking at the screen to looking behind him/her in order to perform a head turn (in response to name calling or social referencing for example). To detect head turns and distinguish between a head turn and just an occlusion of the face, we tracked yaw pose changes and defined two rules: to initiate a head turn the pose had to go from a frontal to one extreme head pose position (left or right); to complete a head turn the pose then had to come back from the same extreme position to a frontal position. More formally, to initiate a head turn the yaw pose had to change from a frontal position $\theta_{yaw} \in \{-20^\circ, +20^\circ\}$ to an extreme $|\theta_{yaw}| > 35^\circ$ within a half-second window. Then to complete a head turn, the yaw pose had to change from the same extreme position back to a frontal position, $|\theta_{yaw}| \leq 20^\circ$, within a half-second window. These time intervals represent a child performing a quick head turn (for example responding to a name-call prompt or social referencing).
An example of this procedure is shown in Figure 3.2.

**Pose-invariant emotion classification**

To analyze children’s behavior in such an unconstrained setting, there is a need for an emotion classification method that can handle faces across varying poses. We employ a modified version of the robust pose-invariant method described in Chapter 2, where we first learn a cross-modality and pose-invariant dictionary. This learned dictionary creates a mapping between facial information from both 2D and 3D modalities and is then able to infer discriminative facial information even when only 2D facial information is available at testing time. For training we use data from Binghamton University 3D Facial Expression database (Yin et al. (2006)), synthesize face images with varying poses, and extract local binary patterns (LBP, Ojala et al. (2002)) and the distances between a subset of facial landmarks as features to learn the cross-modality and pose-invariant dictionary. Using the inferred discriminative 3D and frontal 2D facial features, we train a multi-class support vector machine (Chang and Lin (2011)) to classify four different facial expressions (angry, sad, happy, and neutral) provided by the standard Cohn-Kanade database (CK+, Lucey et al. (2010)). This method focuses on pose-invariant facial expression recognition in images, and outperformed previous state-of-the-art methods.

3.4.2 Human coding

Expert human raters coded facial expressions of emotion, head turns, and instances when a name-call prompt occurred. To code emotion, expert human raters were

<table>
<thead>
<tr>
<th>Table 3.2: Number of video recordings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group</strong></td>
</tr>
<tr>
<td>non-ASD</td>
</tr>
<tr>
<td>ASD</td>
</tr>
</tbody>
</table>
trained to detect facial action units (AUs) from the baby Facial Action Coding system (Baby - FACs, Oster (2003)). They coded ‘Happy’ emotion when activation of the zygomaticus major (AU12) occurred and ‘Other’ otherwise. Expert human raters coded ‘Not Visible’ when the child’s face was covered, out of the field of view, or when more than half of the face was not visible due to head turning away from the camera. Head turning was also coded when a child turned his/her head to look at the practitioner or caregiver (since they were located behind the child). Raters were not blind to diagnostic group, but were blind to stimuli and videos were muted during coding to prevent the influence of vocalizations on the coding of emotions.

Additionally, all name-call prompts were coded separately by a practitioner who marked the frames in the video when the child’s name was called. Coding by the expert human raters and practitioner were performed using the Nodlus Observe XT software Version 11.0 (Nodlus (2015)).

3.5 Validation results and analysis

We now present validation results on the automatic coding of engagement, name-calls including detecting the timing of the prompt and the children’s responses, and children’s emotions. We also present examples of uses of the validated methods for general behavioral analysis. Two-way, average, consistency intra-class correlation coefficient (ICC), and confidence intervals (CI) are used to report inter-rater reliability performance, where a score 0.75 and above is considered excellent (Cicchetti (1994); Hallgren (2012)). Precision, recall, and F1 scores are also employed for validation of emotion classification. While precision considers the accuracy of the decisions made by the automated methods, recall considers the fraction of correct decisions made by the automated methods out of all the decisions made by the expert human raters, and F1 is a combined measure of precision and recall.

Across all participants, 99 video recordings were considered (Table 3.2). Two
Table 3.3: Inter-rater reliability results of automatic codings to expert human raters. Reported are ICC (95% CI)

<table>
<thead>
<tr>
<th>Coding</th>
<th>ALL</th>
<th>non-ASD</th>
<th>ASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engagement</td>
<td><strong>0.85</strong> (0.77–0.89)</td>
<td>0.89 (0.81–0.94)</td>
<td>0.81 (0.69–0.90)</td>
</tr>
<tr>
<td>Name-call response</td>
<td><strong>0.84</strong> (0.67–0.91)</td>
<td>0.86 (0.72–0.93)</td>
<td>0.80 (0.66–0.89)</td>
</tr>
<tr>
<td>Total time exhibiting Happy</td>
<td><strong>0.90</strong> (0.85–0.93)</td>
<td>0.89 (0.81–0.94)</td>
<td>0.90 (0.84–0.95)</td>
</tr>
</tbody>
</table>

Expert human raters were first trained on a reliability dataset (separate from analyzed participants) until they reached excellent agreement, ICC > 0.75. Then, a single expert human rater coded all 99 video recordings, while the second rater coded ~20% of them to verify on-going inter-rater reliability. For the validation here reported, we compare codings between the automated methods and the expert human rater who coded all of the video recordings.

3.5.1 Validation results

Engagement detection

Inter-rater reliability performance of detecting the total time the child was engaged was assessed on a per-video basis, Table 3.3. Reliability between the automatic methods and the expert human raters was excellent, achieving an ICC score of 0.85; while the reliability for the sub groups of participants with and without ASD was also excellent, achieving ICC scores of 0.81 and 0.89 respectively. Since our automatic methods code per-frame, we also analyzed accuracy on a per-frame basis. Across the 161,296 coded frames for engagement (~89 minutes), our methods matched with the expert human raters on 90% of them (with most of the differences located at the start or end of a new behavior). Accuracy of coded frames from the sub groups of participants with and without ASD exhibited similar results. Across 74,670 coded frames from videos of participants with ASD, the agreement accuracy was 85%; where across the 86,626 coded frames from videos of participants without ASD, agreement accuracy was 94%.
Name-call prompt and response detection

To fully assess a child’s behavior to a name-call, we first detect the time a name-call happened and then detect whether or not the child turned his/her head to orient to the name-call. The practitioner manually marked every name-call prompt across all participants by marking the frame in the video when the child’s name was called. For all 99 name-call prompts that were marked, the automatic name-call prompt detection was able to detect every name-call by the practitioner. Fitting an exponential model to the data, we saw a mean absolute time difference of just 0.21 seconds (Figure 3.4(a)).

Child responses after name-call prompts were also coded, where a head turn response was coded if the child performed a head turn within 5 seconds of the name being called. Compared to the expert human raters, the automatic head turn detection method correctly identified 84% of the child responses (46 of the 60 head turn responses, see Section 3.6.1 for a discussion on this, and 37 of the 39 non head turn responses). Overall it achieved an excellent reliability with an ICC of 0.84 (Table 3.3), while the reliability of the sub groups of participants with and without ASD was also excellent, achieving 0.80 and 0.86 ICC scores respectively. By fitting an exponential model to the 46 correctly detected head turns, the mean absolute time difference between the automatic method and the expert human raters was 0.22 seconds (Figure 3.4(b)).

Emotion classification

The automatic emotion classification method outputs \( \ell_1 \) normalized probability weights across emotions. To compare it to expert human raters, we grouped the emotions into the categories coded by raters. Namely, ‘Happy’ considers automatic coding of happy (the key emotion the designed movies is expected to elicit) while ‘Other’ considers angry, sad, and neutral. Automatic coding of emotion was done at a much
Figure 3.4: Area plots of absolute time difference between automatic methods and hand labeled data for name-call prompt detection (3.4(a)) and for head turn detection (3.4(b)). Fitted exponential curves are shown in the dotted red lines.
Table 3.4: To validate the automatic emotion classification and compare it with expert human raters, the precision, recall, and F1-scores are reported when considering all video frames

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Frequency</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>0.20</td>
<td>0.81</td>
<td>0.70</td>
<td>0.75</td>
</tr>
<tr>
<td>Other</td>
<td>0.80</td>
<td>0.91</td>
<td>0.95</td>
<td>0.93</td>
</tr>
<tr>
<td>Overall</td>
<td>1</td>
<td>0.89</td>
<td>0.90</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Overall scores are weighted averages based on frequency.

Table 3.5: To validate the automatic emotion classification and compare it with expert human raters, the precision, recall, and F1-scores are reported across groups

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Frequency</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>0.20</td>
<td>0.85</td>
<td>0.68</td>
<td>0.76</td>
</tr>
<tr>
<td>Other</td>
<td>0.80</td>
<td>0.88</td>
<td>0.96</td>
<td>0.92</td>
</tr>
<tr>
<td>Overall</td>
<td>1</td>
<td>0.87</td>
<td>0.90</td>
<td>0.89</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Frequency</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>0.19</td>
<td>0.75</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>Other</td>
<td>0.81</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
</tr>
<tr>
<td>Overall</td>
<td>1</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Overall scores are weighted averages based on frequency.
finer scale than that of the expert human raters, where the automatic coding estimated probability scores for each emotion and treated each frame independently, whereas the human raters assigned each frame only one emotion and marked time points when the given emotion started and ended. To accommodate these differences, half-second filtering of probability scores and max-voting in ± half-second windows were performed on the automatic coding to determine which emotion was dominantly expressed when the face was detected.

Inter-rater reliability performance of quantifying the total time the child was exhibiting Happy was assessed on a per-video basis, where the inter-rater reliability between the automatic methods and expert human raters was excellent, achieving an ICC of 0.90 (Table 3.3). The reliability for the sub groups of participants with and without ASD was also excellent, achieving ICC scores of 0.90 and 0.89, respectively. Performance of the automatic methods was also validated on a per-frame basis. 136,450 frames (~75 minutes) were coded for emotion across all the participants; tables 3.4 and 3.5 show the precision, recall, and F1-scores of the automatic emotion classification method using the expert human raters as ground truth. Overall, the automatic method achieved high precision, recall, and F1 scores: 0.89, 0.90, and 0.89 respectively.

3.5.2 Behavioral analysis

We now present multiple examples of ASD-risk related behaviors that can be extracted using the above presented and validated methods, including responses from name-call, engagement, and emotion.

During name-call responses it is important to not only code if the child oriented to a name-call but also the latency between when the child oriented and when the child’s name was called. Examples of head turn responses to 3 name-call prompts for a non-ASD and ASD participant are shown in Figure 3.5. The non-ASD participant
not only oriented to all 3 name-call prompts, but all head turns exhibited less than half a second latency. The ASD participant on the other hand, only oriented to 1 of the name-call prompts (the second one in this case) and exhibited a large latency of greater than one and a half seconds. A direct extension of this work is presented and further explored in the companion study (Section 4.2, Campbell et al. (2018)), where these automated methods were used on a study considering 104 toddlers (82 non-ASD and 42 ASD). They reported that toddlers with ASD not only tend to orient fewer times as a response to the name-call, but the mean latency to orient was significantly longer than compared to non-ASD (2.02 vs 1.06 seconds). Only automatic frame-rate methods as the one here presented can detect such important differences and potential risk markers.

The engagement coding allows for further quantification of compliance as well as attention span. Using these methods, (Campbell et al. (2018)) also reported that “there was a significant interaction between diagnostic group and age, with the ASD group showing significantly lower amount of time engaged in the task than the comparison group at older ages only” (this is further explored in Section 4.2). With the validated methods, more in-depth head movement dynamics can be extracted beyond head turning. Figure 3.6 shows examples of extracted head movement tracks and cumulative head movement of three participants while watching the Puppets movie stimuli. Each participant demonstrates a distinct movement profile, varying from sitting still (part. 01) to moving a lot throughout the stimuli (part. 03). This is all automatically measured without any additional hardware, such as motion capture devices.

Emotion related responses to scenes in the movie stimuli, designed to elicit such emotions, as well as spontaneous emotions, are also of great importance for ASD related risk behaviors. Although for the validation of emotion coding each frame was assigned a categorical label, the presented methods provide fine-grained, contin-
uous probability scores for each emotion. Figure 3.7 shows examples of participants without and with ASD expressing Happy when the bunny hops in the Bunny movie stimuli. Both participants react by expressing Happy during this segment, but the participant with ASD exhibits less instances of (high probability) Happy. An extension of this work is presented in our at-home companion study (Section 4.3, Egger et al. (2018)) where it was reported that children with high risk for ASD (self-reported or due to high M-CHAT-R/F score) exhibited lower mean percentage of positive emotions (29.9%) compared to children not at risk (35.1%).

3.6 Discussion

We have been developing an integrated paradigm where short and entertaining movies, carefully designed to elicit behaviors (related to ASD or other developmental disorders risks), are presented on a mobile platform while the device’s camera is recording the participant’s responses (Hashemi et al. (2015b); Campbell et al. (2018); Egger et al. (2018)). This work concentrated on the validation of the computer vision components of this paradigm, in particular for ASD risk markers. Contrary to the standards in the computer vision and machine learning related literature, the ground truth used for validation comes from the population under study and the labeling is done by domain experts. The results indicate high agreement between the proposed automatic methods and the expert human coders, and that the automatic methods can be considered to augment behavioral analysis of early risk markers. Furthermore, these validated methods allow for automatic and objective measurements of high granularity and have many potential benefits for future research of risk markers. This scalability and low cost of the paradigm, basically software only due to the ubiquitous presence of mobile devices, opens the door to deployment on longitudinal studies to track behavioral progress and development. The high granularity of the proposed methods can also lead way to refined definitions of risk markers and behav-
Figure 3.5: Examples of responses from a non-ASD and ASD toddler to name-call.
Figure 3.6: Examples of the head movement of three participants during the Puppets stimuli. 3.6(a) shows the head movements of the participants, where the axis represent pixel coordinates in the video recording. The lines are color-coded with respect to time, where the colorbar on the right represents time (seconds) in the movie stimuli. A log-plot of the cumulative head movement for all three participants is shown in 3.6(b). Figure is best viewed in color.
Figure 3.7: Probability scores of expressing Happy for a non-ASD (top) and ASD (bottom) toddler reacting to a scene during the Bunny movie stimuli. Screenshots of the stimuli are shown in the first row; in this scene in the movie the bunny is jumping and then stopping and making noises while moving its ears and nose. The colorbar on the right indicates probability scores of expressing Happy. Figure is best viewed in color.
ioral trajectories. With these now validated automatic computer vision methods, we presented multiple examples of behaviors that can be extracted. For name-call responses, decisions based on the participant’s head turning and the latency to turning after the name has been prompted can be extracted (Figure 3.5). Attention characteristics and head movement dynamics of the participant throughout the movie stimuli can also be accurately quantified (Figure 3.6). Emotional responses, elicited or spontaneous, during specific events can also be automatically captured (Figure 3.7). While these behaviors are known to be relevant for early risk markers in ASD during actual physical interactions with a trained examiner, it remains to be fully verified that they are still present in this setting of watching movie stimuli. With this said, there are indications that ASD/non-ASD differences in response to name-call, attention, and emotion can be elicited and captured from this setting and by automatic methods (Martin et al. (2018); Egger et al. (2018); Campbell et al. (2018)).

3.6.1 Limitations

There are still many challenges when automatically coding behaviors in unconstrained settings, and even though the validation results were strong, we see this as a starting point for the proposed methods. Some of the missed head turn responses during a name-call were due to occlusions of the child’s faces right before or after a head turn happened (e.g., due to the child’s hands occluding his/her face), or from poor image quality when the child’s head is moving too quickly (using high frame rates available on mobile phones will address this). Since we want a simple, easy to administer setup requiring only the integrated camera on the mobile device, and also want to keep the most naturalistic setting possible by not constraining the child, chances of occlusions of the face may persist. Future algorithms should incorporate hand detection, e.g., Simon et al. (2017), to assist in handling these cases.

There were differences for precision and recall scores for facial expression (emo-
tions) classification between the ASD and non-ASD groups, opening room for improvement of the current methods. High precision scores show that the automatic methods accurately made decisions; whereas, the lower recall scores indicate that the methods missed some codings from expert human raters. One possible explanation for the lower recall scores could come from the training data, since the current methods were trained on images of posed expressions, and as such they perform best at the peak of expressions from participants, hence the high precision scores, and may miss the onset and end of expression. This is a less critical problem if the biomarkers relate to the presence or absence of the emotion event for certain period of time, without the need to use the exact (frame accuracy) time length. We should note that human experts also disagree on the start and end of facial expressions. Additionally, the current facial expression methods were trained on thousands of images from adults and based on adult facial codings. With the annotated data from this work, advancements in techniques to automatically code facial expression, and recent understandings of facial dynamics of individuals with ASD (Guha et al. (2018)), new emotion classification methods based on spontaneous emotion recognition of toddlers are currently being considered by our team. This can be added to the current trained system via modern domain adaptation machine learning techniques.

3.7 Conclusion

In this Chapter, we proposed and validated computer vision methods to automatically code behaviors related to early risk markers of ASD. The algorithms are applied to video recordings from the front camera of a mobile device while the child watched movie stimuli designed to elicit such behaviors. In particular, we focused on automatic methods for quantifying engagement, name-call responses, and emotion responses. We validated the automatic methods using manual coding from expert human raters on a diverse population of toddlers with and without ASD. Additionally,
we showed examples of how the proposed methods can further risk marker research with fine-grained quantification of behaviors. The results suggest these low-cost, objective, and automatic methods can be considered to aid behavioral analysis, and can be suited for objective automatic analysis of large and longitudinal studies.
Findings and extensions using computer vision analysis

4.1 Chapter summary

In this Chapter, we discuss findings where the presented computer vision methods were employed in research studies. Since some these studies contain results outside of the scope of this dissertation, we omit them here and focus on the results relevant to the work presented above (please see citations for full study details).

The first two studies revolved around observational behavior analysis from toddlers and young children. In the first study (Campbell et al. (2018)) we employ the presented methods (a mobile application on a tablet with curated movie stimuli and the developed computer vision algorithms) in a clinicin setting on a diverse population of 104 toddlers and detect atypical engagement and name-call responses of toddlers with ASD. While in the second study (Egger et al. (2018)) we employ the developed computer vision algorithms and extend the developed mobile application to be presented on iPhone devices and be available on the iTunes App Store. After a year, over 1,500 families participated in the study, uploading over 4,000 video
recordings of children in natural settings (homes). Automatic analysis of these videos identified significant differences in emotion and attention by age, sex, and autism risk status.

The last study (Chiew et al. (2018)) was focused on a novel investigation into motivated exploration and memory for a real-life, naturalistic environment. Participants were presented with an introductory statement framing exhibit themes in terms of Promotion- or Prevention-oriented goals before exploring an art exhibit with affectively rich content. They returned 24 hours later for recall and spatial memory tests, followed by measures of motivation, personality, and relevant attitude variables. We captured and analyzed facial expression of the participant’s as they were presented with the introductory statement and found that traits interacted with motivational framing context and facial affect (especially surprised) to predict memory outcome.

4.2 Atypical attention in toddlers with autism

As presented in Campbell et al. (2018).

4.2.1 Motivation and objectives

Early behavioral risk markers for ASD include deficits in social attention and social orienting. Studies of children in the first 3 years of life have shown that a failure to orient to name, attend to distress in others, or show interest in other children, distinguish children with ASD from those with typical development and other developmental delays (Dawson et al. (1998, 2004); Werner et al. (2000)). These signs of atypical social development have now been incorporated into screening and diagnostic instruments for ASD. Both the Modified Checklist for Autism in Toddlers, Revised with Follow up (M-CHAT- R/F; Robins et al. (2015)), and the Autism Diagnostic Observation Scale in Toddlers (ADOS-T; Luyster et al. (2010)) use orienting to name as part of their risk assessment; the former is the widely used screening
measure for ASD and the latter is the gold standard diagnostic instrument for ASD in toddlers.

There is evidence that the combination of reduced attention to social stimuli and slower motor movements could amplify differences in social development in toddlers with ASD. Computer vision analysis (CVA) offers a promising tool for detecting and automatically analyzing attentional and motor behavior in toddlers in response to a social stimuli (Hashemi et al. (2014)). The current study aimed to detect atypical attention behaviors and evaluate whether CVA could reliably measure the consistency and latency of orienting to name in toddlers with ASD versus a comparison group of toddlers without ASD.

4.2.2 Methods

104 toddlers 16–31 months old participated in this study. Twenty-two of the toddlers had ASD and eighty-two had typical development or developmental delay. Toddlers watched movie stimuli on a tablet while the built-in camera recorded their face and CVA automatically coded engagement and responses (Hashemi et al. (2015b)). A practitioner stood behind the child and performed name-calls while the toddler watched the movie stimuli. CVA measured participants engagement and response to name-calls.

Differences in behavior were analyzed between the ASD group and the comparison group. Engagement was compared between groups with a linear model with the main effects of group and interaction of age and group as predictors. Proportion of participants turning to name and consistency of orienting across the three name prompts were compared between the ASD and comparison groups with chi-squared tests. To investigate group differences in the latency to orient to name, time-to-event analysis was used with the time the child initiated head movement toward the examiner as the event and right-censoring at 5 seconds. Cox proportional hazards models
were created with all three name calls for each child linked as repeated measures and with age as a co-variate. Hazards ratios were tested against the null hypothesis of equal hazards between groups with the log-rank test. Kaplan-Meier curves of the cumulative events were constructed to visualize proportion of events in each group over time.

4.2.3 Results

Engagement

Toddlers in the comparison group were engaged in the task for a mean of 89% of the time (stdev=9), compared to 76% (stdev=19) for the toddlers with ASD. There was a significant interaction between diagnostic group and age, with the ASD group showing significantly lower amount of time engaged in the task than the comparison group at older ages only (p=0.03; Figure 4.1).

Response to name-calls

In the comparison group, 49 of 82 toddlers oriented at least once during the 3 trials, whereas 10 of 22 toddlers with ASD oriented to name at least once. This did not constitute a statistically significant difference between proportions of participants from each group orienting to name. In analysis of consistency, however, of the 49 comparison participants who oriented to name, 31 (63%) oriented to multiple name calls, whereas of the 12 ASD participants who oriented to name, only 1 (8%) oriented multiple times (p=0.002; Table 4.1, Figure 4.2(a)). In analysis of latency to orient in non-ASD participants who did orient to name, mean latency between name call and initiation of head movement was 1.06 seconds (stdev=0.72). For ASD participants, mean latency was almost twice as long, 2.02 seconds (stdev=1.43), indicating a significantly longer latency to orient in toddlers with ASD (0.96 sec difference). Log-rank test confirmed a statistically significant group difference in cumulative events
(p=0.003; Figure 4.2(b)). In post-hoc analysis, of toddlers in the comparison group who oriented to name, 94% oriented within 2 seconds of the name call prompt.

4.2.4 Discussions and conclusions

Results indicated that significant differences in task engagement as well as the consistency and latency of orienting to name could be detected between toddlers with and without ASD using measures that were captured with computer vision analysis. Although typically-developing and developmentally-delayed toddlers remained engaged with the task across the age range, older toddlers with ASD were less able to maintain engagement. Our analysis approach also revealed a lack of consistency of orienting in ASD, and replicated past findings in toddlers with ASD. Furthermore, we demonstrated that CVA measures of timing reveal a slower response to a social orienting prompt in ASD. Overall, this study demonstrates that automated coding of orienting behavior via CVA detects atypical social behavior in young children with ASD.
Figure 4.1: Predicted means (lines) and 95% confidence intervals (shaded areas) for proportion of time engaged in the task from models covarying for the age by group interaction (p=0.03). ASD group (blue dashed line) showed less time attending at older ages than the comparison group (pink solid line).

Table 4.1: Number of children in each group orienting to name inconsistently (less than twice) vs consistently (twice or more).

<table>
<thead>
<tr>
<th>Group</th>
<th>Oriented less than twice</th>
<th>Oriented twice or more</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autism</td>
<td>21</td>
<td>1</td>
<td>22</td>
</tr>
<tr>
<td>Comparison</td>
<td>51</td>
<td>31</td>
<td>82</td>
</tr>
<tr>
<td>Total</td>
<td>72</td>
<td>32</td>
<td>104</td>
</tr>
</tbody>
</table>
Figure 4.2: Orienting to Name: (A) Proportion of participants orienting to name in comparison group (pink) vs ASD group (autism spectrum disorder) (blue) on any of the name calls (left panel) vs multiple (2 or 3 times, right panel, p=0.002), with error bars showing 95% CI (confidence intervals) of proportions. (B) Kaplan-Meier plots of cumulative events over time with 95% CI (shading) for each group showing slower orienting in the ASD group (blue dashed line).
4.3 Automatic emotion and attention analysis of young children at home

As presented in Egger et al. (2018)

4.3.1 Motivation and objectives

ASD, affecting 1 in 68 children in the US (Christensen et al. (2016)) is the most common childhood neurodevelopmental disorder. While identification and early intervention of ASD are public health priorities, barriers limit families access to evidence-based screening for ASD. In the US, the median age that a child is diagnosed with ASD is 4 (Christensen et al. (2016)), despite the fact that we can reliably diagnose children at 24 months (Lord et al. (2006)). Many children must wait months or even years for evaluation. The situation is significantly worse in low resource countries. With increasing evidence that early intervention significantly improves outcomes (Dawson et al. (2010, 2012)), there is urgency to bridge the gaps between need and access to care, as well as between science and practice.

Our interdisciplinary team came together to develop novel, accessible, and scalable mobile technology tools to bridge these gaps. Here we present data from Autism & Beyond (Autism & Beyond (2015)) an iOS ResearchKit study using iPhones and a novel behavioral assessment and analysis framework to test the feasibility of a digital health and data science approach to the assessment of young childrens emotions and behaviors in their homes.

Our work emerges from the recognition that a major barrier to early, evidence-based identification and treatment for ASD is the lack of scalable tools for objectively assessing young childrens observed behaviors and emotions. While caregiver-reported information about a child is important, it is far from sufficient for a comprehensive understanding of the child. Currently, the gold standard observational tool for ASD
assessment, the Autism Diagnostic Observation Schedule (Lord et al. (2000)) requires administration by trained professionals in clinical settings, is expensive, and is time-consuming. The lack of feasible, affordable, and accessible observational tools impacts timely identification, the capacity to track developmental change and the effectiveness of interventions, and clinicians and caregivers access to evidence-based knowledge of children’s risk for autism and other developmental and mental health challenges.

Tools that rely on professional administration are simply not scalable. We need tools that capture children’s behaviors in their natural environments, including their homes, schools, and community settings, and can track changes over time. The capacity to assess children outside of clinical and research settings, to engage parents and other caregivers in the collection of data, and to reach families who cannot access services or participate in research will provide information about children that is more ecologically valid and culturally representative. Better population health data and engagement with caregivers will also help to increase awareness about the public health needs of children and families. Apple’s open-source ResearchKit framework gave us the opportunity to extend our work beyond the clinical setting to reach caregivers and children in their homes. Our study, Autism & Beyond (Autism & Beyond (2015)), is an iPhone-based app for caregivers and their children who are 12 to 72 months old. Using ResearchKit, an entire study from a user-driven self-consent process to stimuli presentation, data collection and user feedback is integrated within a user-friendly app downloaded from the Apple App store.

The primary aims of the Autism & Beyond study were to test the acceptability and feasibility of conducting an iPhone-based study with caregivers and young children which included collection of caregiver-report and child video behavioral data and to test a new video-based approach for automatically collecting and quantifying young children’s emotions and behaviors in their natural environments. A secondary
aim was to examine associations of the automatically coded emotions and behaviors with age, sex, and autism risk status. Answering these questions is the critical next step toward our goal of building an integrated digital platform to develop, pilot, and deploy mobile tools that use the capabilities of smartphones, carefully designed stimuli, and automated computer vision and machine learning analytics for screening and monitoring of young children's autism risk, and eventually their broad cognitive and social-emotional development, in their homes, as well as clinical settings.

4.3.2 Methods

The app functions as follows (Figure 4.3). The caregiver downloads the app on his/her iPhone. After presenting on-boarding screens describing the study, the app guides the caregiver through an e-Consent process. If the caregiver meets inclusion criteria and provides consent, s/he then completes three to four brief questionnaires and presents four clinically-informed stimuli (short movies) to the child on an iPhone (or iPad). The camera on the device records a video of the child's face as s/he watches the stimuli. Caregivers can choose to upload whole videos of their child or only the facial landmarks extracted using embedded face detection software encoding. The full videos of the child are uploaded to Duke Health servers and then emotions and attention are automatically encoded (Hashemi et al. (2015b)). The movies were revised versions of the ones used in our iPad study (Campbell et al. (2018); Hashemi et al. (2018)).

4.3.3 Results

Feasibility of collecting video data

Over one year, 1,756 families with children ages 12-72 months old participated in the study, completing 5,618 parent-reported surveys and uploading 4,441 videos recorded in the child's natural settings.
Table 4.2: Reported number of movies viewed and whether the videos could be analyzed using our automatic video coding, for the whole study cohort and for the M-CHAT sub-cohort.

<table>
<thead>
<tr>
<th>Child Movie Stimuli</th>
<th>Bubbles</th>
<th>Bunny</th>
<th>Mirror</th>
<th>Toys and Songs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>20 secs</td>
<td>35 secs</td>
<td>30 secs</td>
<td>30 secs</td>
<td></td>
</tr>
<tr>
<td><strong>Whole cohort</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Recordings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1356</td>
<td>1084</td>
<td>1009</td>
<td>992</td>
<td>4441</td>
</tr>
<tr>
<td>% of whole cohort</td>
<td>77.2%</td>
<td>61.7%</td>
<td>57.5%</td>
<td>56.5%</td>
<td>63.2%</td>
</tr>
<tr>
<td>Full Videos</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>876</td>
<td>698</td>
<td>632</td>
<td>619</td>
<td>2825</td>
</tr>
<tr>
<td>% of whole cohort</td>
<td>49.9%</td>
<td>39.7%</td>
<td>36.0%</td>
<td>35.3%</td>
<td>40.2%</td>
</tr>
<tr>
<td>% who selected video upload</td>
<td>77.2%</td>
<td>61.5%</td>
<td>55.7%</td>
<td>54.5%</td>
<td>62.2%</td>
</tr>
<tr>
<td>Usable Video</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>801</td>
<td>618</td>
<td>535</td>
<td>521</td>
<td>2475</td>
</tr>
<tr>
<td>% of uploaded videos</td>
<td>91.4%</td>
<td>88.5%</td>
<td>84.7%</td>
<td>84.2%</td>
<td>87.6%</td>
</tr>
<tr>
<td>Facial Landmarks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>480</td>
<td>386</td>
<td>377</td>
<td>373</td>
<td>1616</td>
</tr>
<tr>
<td>% of whole cohort</td>
<td>27.2%</td>
<td>22.0%</td>
<td>21.5%</td>
<td>21.5%</td>
<td>23.0%</td>
</tr>
<tr>
<td>% landmark upload</td>
<td>77.3%</td>
<td>62.2%</td>
<td>60.7%</td>
<td>60.1%</td>
<td>65.1%</td>
</tr>
<tr>
<td><strong>M-CHAT cohort</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All recordings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>320</td>
<td>257</td>
<td>243</td>
<td>238</td>
<td>1058</td>
</tr>
<tr>
<td>% of M-CHAT cohort</td>
<td>79.0%</td>
<td>63.5%</td>
<td>60.0%</td>
<td>58.8%</td>
<td>65.3%</td>
</tr>
<tr>
<td>Full Videos</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>207</td>
<td>168</td>
<td>153</td>
<td>154</td>
<td>682</td>
</tr>
<tr>
<td>% of M-CHAT cohort</td>
<td>51.1%</td>
<td>41.5%</td>
<td>37.8%</td>
<td>38.0%</td>
<td>42.1%</td>
</tr>
<tr>
<td>Usable Video</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>185</td>
<td>147</td>
<td>126</td>
<td>127</td>
<td>585</td>
</tr>
<tr>
<td>% of M-CHAT cohort</td>
<td>89.4%</td>
<td>87.5%</td>
<td>82.4%</td>
<td>82.5%</td>
<td>85.8%</td>
</tr>
<tr>
<td>Facial Landmarks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>113</td>
<td>89</td>
<td>90</td>
<td>84</td>
<td>376</td>
</tr>
<tr>
<td>% of whole cohort</td>
<td>27.9%</td>
<td>22.0%</td>
<td>22.2%</td>
<td>20.7%</td>
<td>23.2%</td>
</tr>
</tbody>
</table>
We conducted computer vision analyses on the full videos uploaded. Of the 1,135 caregivers who agreed to upload whole videos, 903 (79.6%) watched at least one movie (mean 2.5; stdev 1.6); 508 (44.8%) watched four movies, 105 (9.3%) watched three, 188 (16.6%) watched two movies, 102 (9.0%) watched 1 movie, and 232 (20.4%) watched no movies. Overall, 2,475 (87.6%) of the videos collected could be analyzed using our computer vision algorithms. Table 4.2 provides rates of usable data for each of the videos. There was no significant difference by sex, age, or autism risk status between usable and not usable videos.

*Automatically quantified emotions and attention*

For each video, we automatically quantified the percentage of positive emotion, negative emotion, and neutral expression, as well as the child's attention (Figure 4.4(a)). We examined whether there were significant associations between coded emotion and attention variables and child age, child sex, and Autism Spectrum (AS) risk, for the whole sample and for the M-CHAT sub-sample.

*Child age* - With increased age, children showed a greater percentage of positive emotion, a lower percentage of negative emotions, and no significant differences in the percentage of neutral expression across all videos (Pearson Correlation Coefficients with age: total positive emotion 0.208, p<.0001; total negative emotion -0.262, p<.0001; total neutral emotion 0.047, p=0.17). Figure 4.4(b) illustrates these relationships. The only significant difference of attention by age was found for the bubbles movie with older children (36-72 months) attending a mean of 84.5% of the time and younger children (12-35 months) attending a mean of 78.5% of the time (p=0.02).

*Child sex* - Because of the preponderance of boys in the sample and the higher rate of autism risk in the boys, we did not separately examine the relationship between the video variables and child sex. Interactions with sex and autism risk are noted
below.

*Autism spectrum risk (AS): Whole cohort*

In these models, we adjusted for child age and sex. In the models stratified by sex, age was adjusted.

*Differences in neutral emotions* - For the bubbles, bunny, and mirror stimuli, higher mean percentage of neutral expressions was significantly associated with high AS risk (high vs. low AR: bubbles: 41.2% (SE 1.8%) vs. 35.8% (SE 1.2%), p=0.01; Bunny 39.2% (SE 1.9%) vs. 33.1% (SE 1.3%), p=0.01; mirror: 36.2% (SE 1.8%) vs. 31.4% (SE1.2%), p=0.02). When we stratified by sex and adjusted for age, we found that these differences were significant for boys but not for girls. For the bubbles, bunny, and mirror, boys with high AR had higher percentage of neutral emotion compared to boys with low AR (high vs. low AR: bubbles: 39.8% (SE 1.8%) vs. 32.9% (SE 1.5%), p=0.001; Bunny 38.0% (SE 1.9%) vs. 31.9% (SE 1.6%), p=0.01; mirror: 36.1% (SE 1.7%) vs. 30.5% (SE1.5%), p=0.01). These differences were not found for girls (high vs. low AR: bubbles: 39.7% (SE 3.7%) vs. 39.2% (SE 2.0%), p=0.89; Bunny 40.4% (SE 4.0%) vs. 34.3% (SE 2.1%), p=0.18; mirror: 34.2% (SE 4.1%) vs. 32.6% (SE2.0%), p=0.72).

*Differences in positive emotions* - In the mirror movie, lower mean percentage of positive emotions was significantly associated with high AS risk status (29.9% (SE 1.9%) vs. 35.1% (SE 1.3%), p=0.02) controlling for sex and age. When we stratified by sex adjusted for age, we found that this difference was significant for boys but not girls (high vs. low AR: Boys: 28.2% (SE 1.9%) vs. 34.2% (SE 1.6%), p=0.01; Girls 33.7% (SE 4.4%) vs. 35.7% (SE 2.0%), p=0.68).

*Differences in negative emotions* - In the bunny movie, lower mean percentage of negative emotions was associated with high AR risk (high vs. low AR: 35.9% (SE 1.9%) vs. 40.4% (SE 1.4%), p=0.04). There was no significant difference by sex adjusted by age (high vs. low AR: Boys: 37.4% (SE1.9%) vs 40.6% (SE 1.6%),
Differences in attention - We found no significant differences for attention for any of the movies when controlling for age and sex. When we stratified by sex and adjusted for age, we found that girls with high AR had significantly lower mean percentage of attention than girls with low AR for the bubbles (high vs. low AR: 86.4% (SE 2.8%) vs. 94.3% (SE 1.5%), p=0.01), bunny (83.0% (SE 3.6%) vs. 91.7% (SE 1.9%), p=0.03), and mirror videos (78.8% (SE 3.8%) vs. 90.1% (SE 1.9%), p=0.01). We found no significant differences for boys by AR status (high vs. low AR: across all videos 88.9% (SE 1.2%) vs 89.3% (SE 1.0%), p=0.79).

Autism spectrum risk (AS): M-CHAT sub-cohort

Of the 407 caregivers who completed the M-CHAT, 334 (82.1%) of their children watched at least one movie (Table 4.2). Neither emotion nor attention showed significant differences in the association with low and high risk M-CHAT groups.

Controlling for age and sex, children with a high M-CHAT score (8 or more) showed significantly lower percentage of positive emotion while watching the bubbles movie compared with medium and low risk children (high vs. medium vs. low risk: 15.8% (SE 4.7%) vs. 29.1% (SE 4.3%) vs. 27.9% (SE 2.8%), p=0.05). Children with high M-CHAT scores watching the bubbles movie had an estimated difference of -12.2% in positive emotion compared to low risk children (score 0-2) (pairwise p=0.02). High risk children had an estimated difference of 13.4% positive emotion compared with medium risk children (score 3-7) (pairwise p=0.03). In particular, boys with high M-CHAT scores watching the bubbles movie had an estimated difference of -15.0% positive emotion compared to low risk children (score 0-2) (pairwise p=0.03) and estimated difference of 13.9% neutral emotion compared to low risk children (pairwise p=0.02). In the bunny movie, boys with high M-CHAT scores showed similar differences, with estimated differences in positive emotion -16.7% (pairwise p=0.01) and neutral emotions 14.6% (pairwise p=0.01).
Across all of the videos, children with a high M-CHAT score showed significantly lower percentage positive emotion compared with medium and low risk children (17.2% (SE 3.4%) vs. 28.6% (SE 3.1%) vs. 24.9% (SE 2.0%), p=0.04). We then examined whether the video extracted emotion and attention variables predicted M-CHAT continuous score controlling for age and sex. With increasing M-CHAT score (i.e., increasing ASD risk), the percentage of negative emotion in the toys and songs video increased (p=0.003, F=8.96); the percentage of neutral emotion increased in the bubbles (p=0.0002, F=14.30), bunny (p=0.0036, F=8.74), and mirror videos (p=0.0004, F=13.46); and across all four videos, the percentage of positive emotion decreased (Bubbles: p<.0001, F=19.2; Bunny: p=0.001, F=10.67; Mirror: p<.0001, F=17.82; Songs/Toys: p=0.01, F=6.8).

No difference in attention in relationship to categorical or continuous M-CHAT score was found.

4.3.4 Discussions and conclusions

Autism & Beyond extends previous work in numerous ways, including (1) using a smartphone to collect data in children’s natural environments (e.g., homes), with a full study integrated in one device; (2) engaging caregivers directly in the collection of video data of young children using developmentally-sensitive video stimuli; and (3) conducting a national population study solely through an app available on the Apple App Store. Our results demonstrate that it is feasible to conduct child development research studies with caregivers and young children in their homes using traditional parent-report surveys and a novel video-based behavioral assessment. We found that two thirds of caregivers were willing to upload the full video of their children while a third opted to upload the extracted facial landmarks. As we are developing the computer vision algorithms, full videos give us the opportunity to test and refine these algorithms. Our long-term goal is to be able to conduct the full coding of the
videos on the phone so that the videos of children not need to be uploaded, improving the children's privacy.

We did find relationships between emotions and attention and age, sex, and AR status. To expand on a few, across all of the videos, percentage of positive emotions increased and negative emotions decreased with age. We did also find significant differences in automatically coded emotions by AS risk status controlling for age and sex. Across all of the videos, except the toys/songs stimuli, children in the whole cohort high AR group showed increased neutral emotion compared to children not in the high risk group. This finding is consistent with previous research showing that children with autism have decreased engagement and decreased range of emotion (Brian et al. (2008); Dawson et al. (2004); Zwaigenbaum et al. (2009)). With the mirror stimulus specifically, children in the high AR group showed decreased positive emotion. We cannot say whether this is related to viewing of faces overall or is specific to looking at themselves. We do have a hint that it may be specific to looking at their own image because we do not find any significant differences by autism risk when the children observed the social sections of the songs/toys video. When stratified by sex, we found that the differences in neutral and positive emotion were specific to boys, while the difference in negative emotions was not.

While the tools we describe here are for research, they are the foundation for the development of clinically-informed screening tools to provide caregivers and clinicians accessible, affordable, and scalable tools for early identification and symptoms monitoring of autism and other developmental and mental health challenges. With this study, we have demonstrated the feasibility of this approach: caregivers are willing to participate; the survey and video data they collected themselves in their homes is high quality; we are able to apply computer vision algorithms to the video data and quantify the observed behavior of young children; and the association of the automatically coded behaviors with age and autism risk are consistent with research
Figure 4.3: Autism & Beyond App. While children are watching neuroscience-based and clinically-informed stimuli (i.e., short movies) on the iPhones screen, the iPhones camera records their facial/head behavior which is then analyzed either in the phone or after data is uploaded (upper row in (a)). All needed information integrated is integrated into the app, from e-Consent to questionnaires to the stimuli and the recording and (partial) analysis. Feedback information from the surveys is provided to the parents/caregivers as well. A few screenshots of the app are provided (lower row in (a)), illustrating the careful design to make it not only scientifically and medically relevant but also appealing and family friendly. All children and adults appearing in the app, e.g., to demo or to describe the study, have given consent. Flow diagram of caregivers engagement with the Autism & Beyond study (b).
Figure 4.4: Automatic coding. The algorithm automatically encodes, from detected features, as marked in the figure (a), both the head position and the emotion while the child is watching clinically-informed movies. From these we can infer their attention, social referencing, and emotional response to the stimuli. Relationship between age and emotion from automatic computer vision methods are presented (b).
conducted in traditional research settings.

4.4 Motivational valence alters memory formation without altering exploration of a real-life spatial environment

As presented in Chiew et al. (2018).

4.4.1 Motivation and objectives

Exploration can appear aimless, but it is not purposeless. In a world of limited resources, learning about the environment via open-ended exploration is crucial to an organism's survival. Exploration enables discovery of new potential rewards and likely threats, and is centrally implicated in learning and memory. Yet despite its clear evolutionary necessity, open-ended exploration of a spatial environment is one aspect of motivated behavior that has received relatively little investigative attention. Moreover, the intuitive relationship between exploration and learning obscures a complicated causality; resolving this causality promises insights into both biological and behavioral bases of memory formation.

There can be little doubt that motivated exploration predicts enhanced memory. Experimental evidence has shown enhanced learning during volitional exploration (Voss et al. (2011); Bridge and Voss (2014); Markant et al. (2014); Gureckis and Markant (2012)), along with increased activation in the hippocampus and other medial temporal lobe substrates of memory encoding (Squire and Zola-Morgan (1991); Voss et al. (2017)). Despite these observations, the causality of relationships between exploration and memory remains ambiguous, because novel stimuli that motivate exploration also reliably elicit changes in neuromodulatory brain systems and directly alter memory formation, via effects on neural plasticity. For example, novelty that elicits exploration in experimental settings also elicits dopamine release. In addition to longstanding research implicating midbrain dopamine (DA) in a broad range
of motivated and adaptive behaviors, including exploration in response to novelty (Krebs et al. (2009); Kakade and Dayan (2002); Minassian et al. (2015)).

Interestingly, however, novelty is not an unambiguous stimulus, and exploration of novelty can be modulated by affect and motivational states. Exploration of novel environments resembles behavioral responses to reward: both elicit approach, behavioral activation, and mesolimbic dopaminergic system activity (Ikemoto and Panksepp (1999); Alcaro et al. (2007)). Moreover, it has been proposed that, from an evolutionary perspective, novelty may hold inherent reward value (Krebs et al. (2009); Kakade and Dayan (2002)). However, novelty is not universally attractive or appetitive: for most organisms, exploratory responses to novelty only occur under conditions of expected reward and safety.

The multivalent nature of novelty creates an opportunity to deconfound effects of motivation and exploration on memory formation. To disambiguate these inter-relationships, we manipulated motivational state prior to entering a spatial context, measured exploratory responses to that context and novel stimuli within it, and then examined motivation and exploration as predictors of memory outcomes. We conducted the study in a physical space an art exhibit examining human relationships to the natural environment (entitled Re-Imagining the Environment, Figure 4.5). The gallery was equipped as an experimental space to elicit and quantify motivated exploration of space and multi-valenced art items. This setting permitted us replicate and extend our prior findings from a virtual spatial environment (Murty et al. (2011)). In addition, we used spatial and item memory measures sensitive to hippocampal and medial temporal lobe components of memory function. These measures allowed us to investigate for previously reported effects: namely, that affect (Bisby and N. (2014); Bisby et al. (2016)) and motivational incentive valence (Murty et al. (2011); Wolosin et al. (2012)) have specific, dissociable effects on memory performance and on the medial temporal lobe memory system (Murty and LaBar (2016)).
In sum, the current investigation aimed to disambiguate the relationships between motivational valence, exploratory behavior and memory, while accounting for momentary affect and potential interactions with individual differences in personality and attitudes. In accordance with accounts of dopamine-driven behavioral activation and exploration behaviors (Ikemoto and Panksepp (1999); Duzel et al. (2010)), we predicted that participants in the Promotion condition would explore the art exhibit more than those in the Prevention condition. Given evidence that reward motivation may specifically improve relational or context memory (Murty et al. (2011); Wolosin et al. (2012)), while threat motivation and negative affect do not (Murty et al. (2011); Bisby and N. (2014); Bisby et al. (2016)), we further predicted enhanced spatial memory in Promotion but no significant differences in item memory, as a function of framing condition. Finally, and, to our knowledge, uniquely in the extant literature, we sought to determine whether the influences of motivational valence on memory formation were attributable to, or independent of, changes in exploratory behavior.

4.4.2 Methods

Ninety-eight participants were enrolled (51 female; mean age 32.9 +/- 1.5 years; range 18-71 years). Participants were recruited from the Duke University and Durham community using posted advertisements. Informed consent was obtained from all participants in accordance with human subjects guidelines established by the Institutional Review Board at Duke University Medical Center. Participants received institutionally standard compensation at the rate of approximately $10/hour, with no additional incentive for performance. Fifty-two participants took part in the Promotion condition and forty-six participants took part in the Prevention condition. Due to technical issues, certain portions of data were missing or unusable. In particular, 16 participants did not have usable exploration video data, and a separate
15 participants did not have facial expression video data. The experiment exhibit is shown in Figure 4.5 and the experimental timeline is shown in Figure 4.6.

Participants were taken to the art exhibit in the experiment gallery (see A in Figure 4.5). All participants entered the gallery space alone, explored, and exited it at will. Prior to exhibit entry, participants were instructed to read in full the exhibit statement, presented on a flat screen monitor at entry (see C in Figure 4.5) and to freely explore the exhibit. Then 24 hours later, they returned for recall and spatial memory tests, followed by measures of motivation, personality, and relevant attitude variables.

Participants facial expressions were recorded as they read the exhibit statement at entry using a high-definition personal camera (GoPro Inc., San Mateo, CA) mounted above the statement display. Facial expressions were automatically classified into angry, happy, sad, surprised or neutral for each frame in the video (Hashemi et al. (2015a)). From the classifiable data, we then examined whether proportions of video frames with a given expressed emotion differed by motivational framing condition (Promotion vs. Prevention) using a mixed-effects logistic regression. Facial expressions were also examined as a predictor of subsequent memory performance. Prior work from our laboratory has demonstrated that individual variability in arousal interacted with motivational context to predict spatial memory, with arousal inversely predicting memory performance under reward but not penalty incentive (Murty et al. (2011)). We investigated whether similar relationships were present in the current dataset by correlating expressed surprise (a putative measure of arousal) with measures of subsequent exploration and memory, separately for Promotion and Prevention conditions.
4.4.3 Results

Results are organized to address the three levels of relationships among motivational state, exploration, and memory: 1) group-level analyses of affect, exploration and memory for the entire exhibit, 2) analyses on the individual item level; 3) analyses of how the motivational framing manipulations interacted with individual beliefs and temperament, including facial expressions of affect during motivational statement reading, to predict exploration and memory. For this dissertation, only a subset of results for 1) and 3) are presented.

Did Affective Facial Responses While Reading Framing Statement Differ with Motivation Condition?

On average, participants in each condition viewed the cue statement for 30 seconds (Promotion M(43) = 30.97 seconds, stdev = 13.41; Prevention M(39) = 30.77 seconds, stdev = 13.94); viewing time did not significantly differ between conditions. Video data of participants facial expressions (N=83) during statement reading was automatically classified as angry, sad, surprised, neutral, happy, or unclassifiable; 20.2% of the data was unclassifiable due to obscured view. Of the classifiable data, across Promotion/Prevention conditions, faces were classified most as neutral (61.1%), then surprised (23.8%) and angry (14.2%), with very few frames classified as happy (0.9%) or sad (<0.01%) (shown in Figure 4.7 separately for each framing condition). Mixed-effects logistic regression revealed that the effect of framing condition was significant for the contrast of surprise vs. neutral expressions [ = -1.3872, SE = 0.5561, z = -2.495, p = .0126], with greater neutral in Promotion vs. Prevention, and greater surprise in Prevention vs. Promotion. No other contrasts reached significance.
Did Affective Facial Expressions While Reading Framing Statement Predict Subsequent Behavior?

Taking expressed surprise as a putative measure of arousal, we measured the proportion of video frames during statement reading where participants facial expressions of affect were identified as surprise. We conducted Pearson correlations between surprise and behavioral measures (total exploration time, item recall success, free recall time, spatial memory accuracy), separately for Promotion and Prevention. Surprise and item recall success (shown in Figure 4.8(a)) were significantly negatively correlated in Promotion \( r(40) = -0.340, p = 0.032 \) but not Prevention \( r(35) = -0.219, p = 0.206 \); however, these correlations did not significantly differ in strength \( z = -0.54, p = 0.589 \), two-tailed). Surprise and spatial memory accuracy (shown in Figure 4.8(b)) were significantly negatively correlated in both Promotion \( r(41) = -0.708, p < 0.001 \) and Prevention \( r(40) = -0.374, p = 0.017 \); this correlation was significantly stronger in the Promotion group \( z = -0.212, p = 0.034 \), two-tailed). Finally, the correlation of surprise with spatial memory accuracy was significantly stronger than with item recall success \( z = -0.212, p = 0.022 \), two-tailed). While these correlations of surprise with behavioral measures should be considered exploratory, the negative correlation between surprise and spatial memory in the Promotion condition was robust, surviving Bonferroni correction for multiple comparisons. In sum, in the Promotion condition, the greater the surprise (i.e., arousal) elicited by the motivation manipulation, the poorer subsequent memory was, particularly spatial memory.

4.4.4 Discussions and conclusions

The present study compared profiles of volitional exploratory behavior under promotion and prevention motivation in a complex, real-life spatial environment, employing multiple memory measures characterizing both item and relational memory, to examine exploratory encoding behavior as a potential mechanism for motivated memory.
Further, we explicitly examined the role of individual difference measures and their potential interactions with motivational context to predict encoding behavior and memory outcomes.

The prediction that participants would show greater exploration and correspondingly enhanced contextual memory in the Promotion vs. Prevention condition was not fulfilled in the present data, at least in terms of exploration time and measures of recall and spatial memory. Rather, we observed that exploration time and engagement during encoding were more tightly correlated to subsequent memory in the Promotion condition, suggesting that the Prevention manipulation disrupted typical depth-of-encoding relationships. Additionally, surprise expressed in response to the motivational manipulation was negatively associated with subsequent spatial memory, specifically in the Promotion condition. Finally, individual differences in personality and attitude variables predicted exploration and memory outcomes; regression analysis indicated both main effects of individual differences, and interactions with motivational context, on these outcome variables.

**Stronger Influence of Surprise on Memory in Promotion Condition Parallels Our Previous Arousal Findings**

Facial expressions during statement reading and subsequent behavior varied with framing condition. Participants expressed more surprise under Prevention than Promotion. Further, surprise was negatively associated with spatial memory in both conditions, but this relationship was stronger in the Promotion condition. Given that surprise is associated with heightened arousal, relative to a neutral emotion state (Collet et al. (1997)), the inverse association between surprise and spatial memory can potentially be liked to prior findings from our laboratory (Murty et al. (2011)), where high arousal predicted poorer spatial memory, specifically under reward. Effects of surprise on memory encoding have been mixed in the literature: surprising
events can disrupt cognitive processing (Garrido et al. (2011); Schroger and Wolff (1998)), but may also signify potential reward predictors during goal pursuit; enhanced memory has been observed for task- incidental, surprising stimuli encountered during reward anticipation (Murty and Adcock (2013)). In the present results, surprise appeared to have an impairing effect: memory for the exhibit space was impaired, and no enhancement in memory was observed for the exhibit statement itself, as a function of surprise. Only a minority of subjects mentioned the statement during recall (34 of 91 subjects with usable free recall data); but given that the exhibit statement was not obviously an art piece in the exhibit, it is possible that it was not considered test memoranda. A forced-response recognition memory paradigm, would have allowed direct evaluation of memory for the statement itself. Although memory for the surprising statement itself was not definitively assessed, these results add to a mixed literature regarding surprise effects on memory, indicating that surprise may disrupt memory for subsequent events.

Overall, the present study provided a novel investigation into motivated exploration and memory for a real-life, naturalistic environment. We observed that motivational framing did not affect overall motivation to remain in the novel spatial context, but instead altered the relationship between encoding behavior and memory outcomes. Although increased exploratory behavior is one mechanism of improved subsequent memory performance linked to hippocampal function (Voss et al. (2011); Markant et al. (2014); Voss et al. (2017)), the current findings suggest that motivational contexts elicit mechanisms that constrain memory performance independently of effects on exploration or encoding, at least in terms of exploration time. Additionally, individual differences in personality, attitudes, and affective response interacted with motivational context to improve predictions of behavior. Given our stimuli, these findings also help characterize predictors of motivated engagement with and memory for sustainability-related information. By providing a characterization of
multiple, interactive influences on memory in a naturalistic environment, the present
data offer additional insights into mechanisms for further investigation and an ac-
count that more closely parallels how motivated memory unfolds during daily life in
a complex world.
Figure 4.5: (A) Schematic of exhibit space (13.1m x 6.25m; 82 square metres). A partial wall occluded the space at entry and displayed a monitor with the Promotion or Prevention-themed exhibit statement. (B) Examples of artwork in the exhibit, which explored the relationship between humans and the natural world. Eight pieces of art, of different media, were displayed. (C) Promotion and Prevention versions of the exhibit statement, where human response to environmental change was framed as pursuit of desired outcomes (Promotion), versus prevention of undesired outcomes (Prevention), to elicit distinct motivational states, as indexed by facial expressions of affect.

Figure 4.6: On Day 1, participants read either a Promotion or Prevention-oriented statement at entry and then freely explored the exhibit space, ending their visit at will. A wall-mounted camera recorded participants as they read the statement and an automatically classified facial expression as affective responses to the manipulation. A ceiling-mounted video system recorded participant activity through the exhibit: these data were used to calculate exploration time. Twenty-four hours later, participants provided open-ended free recall of their visit. Participants next completed a spatial memory test of the exhibit, followed by individual difference measures.
Figure 4.7: Participants expressed significantly more surprise (and correspondingly, less neutral expression) in Prevention vs. Promotion. These findings confirm that participants had differing affective responses to the Promotion and Prevention-oriented versions of the exhibit: specifically, participants expressed more surprise in response to the statement in the Prevention condition.

Figure 4.8: Item recall success and spatial memory accuracy as a function of expressed surprise and motivational framing. Expressed surprise (a) negatively predicted subsequent item recall success in Promotion framing (n.s. in Prevention framing); and (b) negatively predicted subsequent spatial memory accuracy in both framing conditions. This relationship was significantly stronger in Promotion framing. Line shading indicates standard error.
5.1 Chapter summary

In this Chapter, we present a framework to calibrate a pre-trained deep network to deployment domains where limited training information is available. In many real-world scenarios, deep networks that achieve state-of-the-art results on benchmarks are still prone to suffer performance degradation when deployed due to shifts between the training and deployment domains. This chapter focuses on a framework to calibrate pre-trained face models to specific deployment domains. In this Chapter, face synthesis approaches are proposed and then optimally selected both at training and testing, leading to more efficient training and improved performance at testing without having to re-design a new network.

5.2 Introduction and related work

Recent advances in DNNs have greatly impacted the face recognition community. DNNs that have achieved state-of-the-art performances on benchmark datasets trained
using millions and hundreds of millions of training images (Parkhi et al. (2015); Schroff et al. (2015); Taigman et al. (2015)). With this said, they are still prone to suffer severe performance issues when deployed in many real-world scenarios. These performance issues stem from shifts between the training and deployment domains such as image resolution, lighting conditions, occlusions, ethnicities of subjects, among others. For example, consider the deployment scenario with a face recognition system at a remote checkpoint location, where the task is to identify images of subjects captured on a mobile phone. Following most state-of-the-art methods, in particular those used by systems without access to the hundreds of millions of private collections, the training set will be made up of labelled images taken from the internet (usually of celebrities with makeup and at ideal lighting conditions). While in this deployment scenario, the testing set contains only images taken from the user’s mobile phone, possibly at poor lightning conditions and reduced image quality.

One straightforward way to address this scenario is to collect a sufficient amount of labelled data in the expected deployment setting; however, this is an intensive and tedious solution and is often times impractical. A more practical solution is to obtain a limited set of labelled data in the deployment setting and then use it to calibrate (adapt) a pre-trained model. A major concern with using limited training data is overfitting. Fortunately, with advances in computer vision and graphics, it is possible to enrich a limited dataset with face synthesis methods where one uses the provided annotated face images to generate new ones (Hu et al. (2016); Masi et al. (2016)). For face recognition, intra-class face synthesis approaches (detailed in Section 5.3) are especially powerful since they enrich the training dataset while at the same time retain the original class labels. These face synthesis approaches can also be exploited at testing time to generate more realizations of testing images.

Although face synthesis approaches have shown to be powerful methods to enrich
limited datasets, the current driving force behind synthesizing face images is to just generate as much data as possible to feed into a DNN. Performing face synthesis in such a blind manner is not only very time consuming and computationally expensive during training, but can be prone to generating redundant data that will have very little impact or generating data that is too extreme and will harm the calibration method. Smart synthesis during deployment (testing) is of course even more critical.

Enriching datasets with synthesized data has been explored in many visual recognition tasks including body pose estimation (Rogez and Schmid (2016)), text localization (Gupta et al. (2016)), gaze estimation (Shrivastava et al. (2016)), and face recognition (Hu et al. (2016); Masi et al. (2016)). Notably, Hu et al. (2016) employ face-swapping to clone combinations of 6 facial regions. This enriches a face dataset by generating both novel intra-class and inter-class (new subjects) face images. The work in Masi et al. (2016) fits a generic 3D morphable model (3DMM) to face images and enriches face datasets by synthesizing with different shape, pose, and mouth variations. At testing they perform synthesis on the test images but only to create pose variations. These works focused on training DNNs from scratch, without any intelligent selection of the synthesis, and demonstrated that these face synthesis methods improved performance. We will exploit these synthesis methods.

The work in Paulin et al. (2014) explored the effectiveness of simple image transformations (scaling, mirroring, flipping, etc.) for the task of image classification. They demonstrated the importance of iteratively selecting transformation types that are most informative during training to increase efficiency. In addition, they showed the importance of utilizing informative transformations at both training and testing. At each iteration for selecting a transformation type, their approach trains classification models for every possible transformation type, then picking the type that provides the greatest increase in performance. Training models for every transformation type at each iteration can be computationally expensive, and does not easily
allow for combinations of transformations to be considered.

In this chapter, we explore the effectiveness of face synthesis for calibrating a pre-trained DNN to multiple deployment scenarios where limited training data is available. To achieve this we develop face synthesis methods and propose information-driven approaches to exploit and optimally select face synthesis methods and samples. We show that our approaches lead to more efficient training and improved testing performances.

In this chapter, our main contributions are:

• We develop a Poisson face-swapping method and person-specific 3D morphable model to synthesize novel face images.

• We propose information-driven approaches to guide face synthesis during training and testing;

• We demonstrate how to efficiently use face synthesis to calibrate a DNN given multiple deployment scenarios, thereby demonstrating how DNNs can become even more powerful than what they were designed/trained for.

5.3 Face synthesis methods

Given a limited training dataset from the target domain, the goal is to be able to enrich the dataset while preserving the annotated class labels of the target domain; in this section we outline three such face synthesis methods. First, we describe our Poisson face swapping method that couples face-swapping with Poisson editing to synthesize realistic face images. In addition, we use a 3DMM to generate novel realizations and viewpoints of a face image through deforming a fitted person-specific face model and capturing realizations at different head poses. We preprocess each face image by first detecting facial landmarks (King (2009)), and then using 7 landmarks locations (namely the two right eye corners, two left eye corners, nose tip, and
the two mouth corners) we align the face to a canonical frontal face model through a similarly transform. After alignment, we finally combine the two methods.

For notation, images provided by the target dataset are referred to as base images. $B\{n, m\}$ represents a base image from subject $n$ and image number $m$, where $n \in N$ and $m \in M$. $S(B\{n, m\})$ represents a set of synthesized face images generated from the base image $B\{n, m\}$, where $|S(B\{n, m\})|$ is the total number of face synthesis types, and each synthesis type $i \in |S(B\{n, m\})|$ is represented by $s_i(B\{n, m\})$.

5.3.1 Poisson face-swapping

Given any two face images, we synthesize new face images through combinations of their face regions. Motivated by Hu et al. (2016), we use automatically detected facial landmarks to define 3 face regions: eyes, nose and mouth, and rest of the face (see Figure 5.1). To synthesize a new face image, we define the triplet $(b, d, c)$ where $b$ and $d$ correspond to two images that will be mixed, and the bitcode $c \in \{0, 1\}^3$ defines which face regions will be taken from each image. A zero in the bitcode $c$ represents the corresponding face region will be taken from image $b$, whereas a one represents the corresponding face region will be taken from image $d$.

Just swapping the face regions introduces unnatural artifacts, including major image gradients around the swapped face regions and drastic contrast differences between the swapped regions. More realistic synthesized face images can be generated by viewing the process of face swapping as an application of guided interpolation that can be solved via Poisson image editing methods (Di Martino et al. (2016); Pérez et al. (2003)). Guided interpolation aims at seamlessly cloning novel objects or image sections into a background image. In this case of face swapping, the objective is to seamlessly clone face regions of one face image over to another face image. The background image is created based on the bitcode $c$ and is defined as the image created from the combination of the ‘rest of the face’ region and other face regions.
that share the same code entry as the ‘rest’ entry. Let $\mathcal{B}$ be the background image and $\Omega \in \mathcal{B}$ be the domain of the face regions which we wish to replace, where $\partial \Omega$ is the boundary of these face regions. The known image values in $\mathcal{B}$ are denoted as $f^*$, while $f$ are the unknown image values defined over the interior of $\Omega$. Furthermore,
let $g^*$ be the known image values of the face regions we wish to clone onto the face regions in the background image, where its gradients $\nabla g^*$ are used as the guidance vector field ($\nabla = \left[\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right]$ is the gradient operator). The goal is to minimize the difference of the gradient vector fields between the background image and the desired face regions we wish to clone,

$$\min_{f} \int_{\Omega} \| \nabla f - \nabla g^* \|^2, \quad s.t. \ f|_{\partial \Omega} = f^*|_{\partial \Omega},$$

whose solution is the Poisson Equation over the domain $\Omega$ with imposed Dirichlet boundary conditions,

$$\nabla^2 f = \nabla^2 g \text{ over } \Omega, \text{ and } f|_{\partial \Omega} = f^*|_{\partial \Omega},$$

where $\nabla^2 = \left[\frac{\partial^2}{\partial x^2}, \frac{\partial^2}{\partial y^2}\right]$. Many approaches can be used to solve Equation (5.2), for this work we use the finite difference method implementation (Di Martino et al. (2016); Pérez et al. (2003)). Thus the final synthesized image will contain computed pixels $f$ inside the swapped facial regions defined by $\Omega$, and background pixel values $f^*$ outside $\Omega$. Figure 5.1 illustrates examples of our proposed face-swapping method, which is both very simple (based on swapping) and very realistic (thanks to Poisson editing).

5.3.2 Person-specific face morphisms and poses

To generate novel morphisms of any face image, we first iteratively morph a 3DMM so that corresponding landmarks between the face image and 3DMM align with one another. We utilize the Basel Face Model (Paysan et al. (2009)), a linear principal component analysis 3DMM parameterized by 199 shape principal components. It is a very dense model consisting of 53,490 depth vertices and 106,466 faces. For ease of computation, we concatenate the depth vertices and there for consider $53,490 \times 3 = 160,470$ vertices. From a 3DMM new 3D face models, $\text{BFM}(\alpha) \in \mathbb{R}^{160,470}$, are
Figure 5.2: Organization of collections for base and synthesized face images. For any image \( m \) from subject \( n \), \( B\{n,m\} \) represents the base face image and \( S(B\{n,m\}) \) represents the set of synthesized face images. Face synthesis types, \( s_i \), span the column space of \( S \).

synthesized via

\[
BFM(\alpha) = \mu + U\alpha,
\]  
(5.3)

where \( \mu \in \mathbb{R}^{160,470} \) is the mean face shape, \( \alpha \in \mathbb{R}^{199} \) contains the shape parameters, and \( U \in \mathbb{R}^{160,470 \times 199} \) is the shape basis.

For learning a person-specific face model from an input face image, the goal is to determine the optimal \( \alpha \) values that correctly register the BFM model to the input face image. Let us introduce two rigid transformation parameters, a scale and
rotation parameter $R$ and a translation parameter $t$. Then the following two term loss function is minimized:

$$E(\theta) = E_l(\theta) + \eta E_s(\theta),$$  \hspace{1cm} (5.4)$$

where $\theta = \{\alpha, R, t\}$ contains the shape and rigid transform parameters. The two terms represent a landmark term $E_l$ and a regularization term $E_s$, where the regularization term is weighted by the stiffness parameter $\eta$. More specifically these terms are given by

$$E_l(\theta) = \sum_{(p,l)\in L} \|R (\mu_p + U_p\alpha) + t - l\|^2,$$  \hspace{1cm} (5.5)$$

$$E_s(\theta) = \|\alpha\|^2,$$  \hspace{1cm} (5.6)$$

where $\mu_p$ and $U_p$ represent the rows corresponding to vertex $p$ in $\mu$ and $U$ respectively. The term (5.5) minimizes the distances between the corresponding BFM and image landmarks $L$. The regularization term (5.6) enforces small values for the shape parameters $\alpha$ and is guided by the stiffness parameter $\eta$. The proposed method is an adaptation of the method in Mora and Odobez (2012), where the authors focused on fitting a 3DMM to an input face mesh.

The fitting defined by (5.4) is an iterative process where the stiffness parameter $\eta$ is increased after each iteration. As $\eta$ increases, it restricts the amount the 3DMM can deform. Convergence is achieved when the parameter set differs by less than a small margin between iterations. Once the optimal face model $\text{BFM}^*$ has converged, each vertex is assigned a texture index by directly sampling from the nearest pixel location on the input image.

Since the learned person-specific face model $\text{BFM}^*$ is derived from a 3DMM, new face models $Z(\tilde{\alpha})$ based on the input face image can be synthesized by varying the shape parameter $\tilde{\alpha}$,

$$Z(\tilde{\alpha}) = \text{BFM}^* + U\tilde{\alpha}.$$  \hspace{1cm} (5.7)$$
Notice that (5.7) and (5.3) are identical except the mean face shape in (5.7) is replaced by the learned person-specific face model. Since $Z$ and $BFM^*$ have one-to-one vertex and face correspondences, the texture from the learned person-specific model can be directly transferred to the synthesized face model $Z$.

Furthermore, we can generate novel poses of any 3DMM by rendering at different viewpoints. Examples of face morphisms and pose variations are shown in Figure 5.1(d).

5.3.3 Combinations of methods

Face synthesis can be performed through combinations of the methods described above. For the work presented here we focus on intra-class synthesis, where the synthesized face images are always assigned a class label belonging to the known training classes. We refer the reader to Hu et al. (2016) for work where face-swapping was performed to create unseen labels. If the training dataset contains $N \cdot M$ total images where $N$ is the total number of subjects and $M$ denotes the images per subject, the total data enrichment increases exponentially with $M$. Specifically the enriched dataset contains $M(1 + |c|! \frac{M!}{M-2!(M-2)!}) |\text{morphisms}| |\text{poses}|$ synthesis types per subject.

5.4 Information driven synthesis for calibration

Face synthesis is known to provide meaningful information for the task of face recognition (Masi et al. (2016)); however, the process of synthesizing face images is very time consuming, requires large storage space, and can potentially bias training. Often, many of these synthesized faces contain redundant information across the other synthesized faces and the base images they are produced from. We propose two information theoretic driven approaches to make face synthesis more efficient for training and increase performance at testing.
5.4.1 Model calibration

To calibrate the deep features to a given deployment scenario, we first normalize the deep features via $L_2$ normalization, then freeze the network except for a newly added fully connected layer. This added layer serves as an embedding to map the deep features to the given deployment scenario. There are many different embedding constraints that can be deployed, such as pair-wise (Davis et al. (2007)) or low-rank (Qiu and Sapiro (2015)). For this work we use the triplet-loss (Schroff et al. (2015); Weinberger and Saul (2009)). Let $x(I)$ denote the $L_2$-normalized feature representation of face image $I$, where $\|x(I)\| = 1$. The triplet loss handles a triplet of examples, namely $\{x(I^a), x(I^+), x(I^-)\}$, where $x(I^+)$ is the positive feature representation sharing the same class as the anchor $x(I^a)$, whereas $x(I^-)$ is the negative representation belonging to a different class. The goal is to learn the embedding $W$, so that the embedded feature space $\phi(x) = Wx$ minimizes the distance between the anchor-positive pairs while maximizing the distance between the anchor-negative pairs. For notation sake, the sub and super-scripts of $I$ will be transferred to $x$; so $x$ represents the feature of any image input $I$, and furthermore, $x^a, x^+, x^-$ represent the normalized feature representations of anchor, positive, and negative images respectively. The triplet loss to be minimized is as follows:

$$E_{\text{triplet}} (\phi(x^a), \phi(x^+), \phi(x^-)) =$$

$$\max\{0, \gamma + \|\phi(x^a) - \phi(x^+)\| - \|\phi(x^a) - \phi(x^-)\|\}.$$  

Here $\gamma$ acts as a margin parameter between the negative and positive pairs, The task of choosing a triplet is crucial. Similar to Parkhi et al. (2015); Schroff et al. (2015), we perform hard-negative mining for triplet selection. For each class, we sample an anchor-positive pair, and extend each pair by randomly sampling a negative sample. If the distance between the anchor-negative pair is greater than the distance between the anchor-positive pair, then we do not use the negative sample to update; however,
Figure 5.3: Visual results for optimal synthesis selections for training and testing. (a) Images provided by the dataset. The rows in (b) show results for optimal selection of synthesis types for training guided by ME (top) and MMI (bottom). ME favors synthesis at extreme poses and morphisms, whereas MMI chooses a more balanced synthesis subset both in training and in testing. The columns in (c) shows synthesis selections guided by MMI for testing from the right-most base images in (a).

we still update with the anchor-positive pair. An epoch considers all anchor-positive pairs in the training set. We use an initial learning rate of 0.1 and convergence is achieved when the embedding differs by a small margin between epochs (usually 50 epochs).

5.4.2 Modeling face synthesis as a Gaussian Process

Let the features from a face image generated from synthesis type \( i \) be represented by \( L_2 \) normalized \( d \)-dimensional vector \( s_i(B\{n, m\}) \in \mathbb{R}^d \), where a synthesis type is
defined as any combination of the face synthesis methods in Section 5.3. Features from synthesized face images with redundant information will have high similarities with one another. We model the face images as a Gaussian Process (GP) which is represented by a mean function and a positive-definite kernel function $K$. Furthermore, we combine all synthesis outputs across the subjects to create a collection $S$ of face images from different types of synthesis where each column $s_i \in \mathbb{R}^{dMN}$ represents a different synthesis type and the rows are organized in subject-specific blocks (Figure 5.2(b)). From $S$, we define the kernel function across each pair $(i, j)$ of synthesis types as $K_{s_i, s_j} = \text{sim}(s_i, s_j)$, where $(s_i, s_j) \in \vert S \vert$ and $\text{sim} (\cdot)$ is the cosine similarity. A GP allows us to model any synthesis type as a Gaussian distribution whose conditional variance is given by $\sigma_{s_i|S} = K_{s_i, s_j} - K_{s_i, S} K_{S, S}^{-1} K_{S, s_i}$ where $K_{s_i, S}$ is the covariance vector between $s_i$ and $S$. For computational efficiency it is beneficial to use a kernel function with compact support. One way is to set a small threshold $\epsilon$ where one removes all synthesis types $i$ for which $\vert K_{s_i, s_j} \vert < \epsilon$.

Our objective is to select an optimal subset of synthesis types that are most representative to space of all possible types. A straightforward approach is to minimize the conditional entropy $H(\cdot)$ of the non-selected synthesis types given the already selected subset $S^*$:

$$\arg \min_{S^*} H(S \setminus S^* | S^*) \Rightarrow \arg \max_{S^*} H(S^*).$$

(5.8)

In turn this selects the optimal subset with maximum entropy, and in practice it biases extreme synthesis types (Figure 5.3(b)), which is not ideal. These shortcomings of optimal selection using entropy criterion have also been observed in Krause et al. (2008); Qiu et al. (2011) for tasks related to sensor placement and action attribute learning. We will address this problem next.
5.4.3 Optimal synthesis selection for training

To diminish the bias of selecting extreme synthesis types, we add the constraint of selecting synthesis types that are most representative of all the non-selected types. Specifically, we want

\[
\arg \max_{S^*} H(S \setminus S^*) - H(S \setminus S^* | S^*)
\]

\[
\Rightarrow \arg \max_{S^*} I(S^*; S \setminus S^*),
\]

which is equivalent to selecting the subset that maximizes the mutual information \(I(\cdot)\) between the selected types \(S^*\) and the non-selected types \(S \setminus S^*\).

Selecting the optimal subset can be done with a greedy algorithm as follows. Initialize \(S^* = \emptyset\). Then iteratively choose the next best synthesis type \(y^*\) from \(S \setminus S^*\) that provides the maximum increase in mutual information,

\[
\arg \max_{y^* \in S \setminus S^*} I(S^* \cup y^*; S \setminus (S^* \cup y^*)) - I(S^*; S \setminus y^*)
\]

\[
\Rightarrow \arg \max_{y^* \in S \setminus S^*} H(y^* | S^*) - H(y^* | S^*),
\]

where \(\bar{S}^*\) denotes \(S \setminus (S^* \cup y^*)\). Since the conditional entropies are from a Gaussian random variable, they have the closed form solution

\[
H(y^* | S^*) = \frac{1}{2} \log(2\pi e \sigma_{y^* | S^*}^2).
\]

This greedy approximation algorithm can be solved in polynomial-time (Krause et al. (2008)).

So far we have focused on a collection of synthesized face images only; however, the optimal selection of \(k_{\text{train}}\) synthesized face images can be guided by incorporating the provided base face images of the subjects, \(B\), simply initializing \(S^* = B\). Then we iteratively choose the next best synthesis type according to (5.10), until \(|S^*| = k_{\text{train}}.\)
5.4.4 Optimal synthesis selection for testing

In the previous section we were concerned with determining the types of face synthesis that provide the most information with respect to all of the training data in order to calibrate (adapt) the model to the target domain. At testing time we wish to further exploit the synthesis types that were performed in training to improve testing performance; however, the trade-off between accuracy and computation is crucial, thus using synthesis methods with redundant information is not desirable. Selecting the synthesis types that are most similar to each other will lead to selecting a local group of synthesized face images with redundant information. Instead we want to select synthesis types that are similar to each other but are also informative with respect to types that have already been selected. Again, we can use the same maximum mutual information approach (5.9), but now further restrict the selections to those that were employed during training. Thus the algorithm for picking $k_{test}$ optimal synthesis types for testing is to first initialize $S^* = B$. Then iteratively choose the next best synthesis type $y^*$ from $S \setminus S^*$ according to (5.10) until $|S^*| = k_{test}$.

At testing it is not guaranteed that multiple of images per subject will be available, thus for testing we restrict the face synthesis collection to be defined by face morphism and pose variations only. Selecting the optimal synthesis types for testing is very efficient since it does not require any additional synthesis to be performed since it only considers synthesis types that are a subset of the synthesis types that were performed during training.

5.5 Experimental validation

We use the state-of-the-art VGG-Face (Parkhi et al. (2015)) DNN to extract feature representations of our face image. To date, VGG-Face\(^2\) is the highest performing

\(^2\) [http://www.robots.ox.ac.uk/~vgg/software/vgg_face/](http://www.robots.ox.ac.uk/~vgg/software/vgg_face/)
publicly available model for facial recognition on the gold-standard *Labeled Faces in the Wild* dataset (Huang et al. (2007)), achieving 98.95% verification accuracy. It is a 16 layer DNN trained on the *VGG Face Dataset* (Parkhi et al. (2015)) which contains 2.6 million images taken from the web of over 2,000 celebrities. A driving component of VGG-Face’s success is due to the access of millions of images taken from the web during training, thus sharing and capturing the same domain as many other validation datasets. For example, on the YouTube Faces (Wolf et al. (2011)) benchmark, which consists of videos taken from YouTube, it also achieves near-perfect accuracy: 97.30%. To validate our proposed calibration approaches, we picked two datasets that are from different domains than the web, and where VGG-Face performs less than optimal on. Namely we chose the OFD\(^1\) and CASIA NIR-VIS 2.0 (Li et al. (2013a)) datasets, where out-of-the-box VGG-Face achieved 80.70% and 67.47% rank-1 classification scores respectively. The OFD dataset contains prominent illumination challenges, whereas the CASIA NIR-VIS 2.0 dataset contains images from different modalities.

Since the last layer of VGG-Face trained on labels from the VGG Face Dataset, we discard it and use the second to last layer as our deep feature representation. We perform principal component analysis to reduce the deep feature representation to 512 dimensions. We then learn a 512-by-512 embedding \(W\) via the triplet-loss to adapt the model to the different dataset domains. Lastly, we use the cosine similarity score to perform matching. If synthesis is performed during testing, the average matching score across all synthesized images from a given testing image is used. Note that the method here proposed will enjoy this state-of-the-art method without having to re-design it when adapting it to new domains.

It takes our system, Intel Core i7 5820K computer with 64GB DDR4 RAM and an NVIDIA GeForce Titan X, 800 milliseconds to synthesize new faces.

\(^1\) [http://gr.xjtu.edu.cn/web/jianyi/tt](http://gr.xjtu.edu.cn/web/jianyi/tt)
5.5.1 Results on OFD

*OFD* is a Chinese face dataset containing 33,669 images across 1,247 subjects. Images from this dataset were taken in a controlled setting, where poses, lighting conditions, background color, and facial accessories were controlled, making it an ideal dataset to demonstrate effectiveness of calibrating a trained model to a new deployment scenario. Without any calibration, *VGG-Face* performs with an 80.70% rank-1 classification score in accordance with our validation procedure explained next.

For all results presented we conduct 5-fold cross-validation. Thus, we first separate the subjects into 5 validation partitions (around 250 subjects per partition). For each validation partition, we test with only the subjects in a given partition and train with subjects in the remaining partitions. We compute classification scores from the data in the testing partition using a gallery depth of 1 and probe depth of 4 per subject (see Figure 5.4). We focus our experiments around scenarios where limited training data is provided: \( M = 2 \) images per subject and \( N = \{5, 10, 20, 50\} \) subjects are provided. Thus we further divide each training partition into subsets with non-overlapping training subjects. For a given validation partition, classification results are computed by averaging the testing partition’s scores across all the subsets in the respective training partition. Then the final results shown are the average scores throughout the 5 validation partitions. We also explore the effects of \( k_{\text{train}} \) and \( k_{\text{test}} \).
Figure 5.5: Rank-1 classification performance as the number of training subjects, \( N \), increases. Initial rank-1 score for un-calibrated model is 80.70\%. Our proposed training and testing approaches incorporating MMI, require less than 5\% of the total amount of synthesized data at training and achieve highest performance. It also outperforms using ME to select optimal synthesis. For these experiments \( k_{\text{train}} = 10 \) and \( k_{\text{test}} = 5 \).
which define the size of the training and testing synthesis subsets respectively.

**Incorporating synthesis at training and testing.** For each subject provided in the training set, we generate a total of 280 synthesized face images defined by a combination of 8 face-swappings, 5 morphisms, and 7 pose variations. Examples of the synthesized face images for a given subject can be seen in the first two rows of Figure 5.2(b). We compare our proposed optimal synthesis selection guided by *maximum mutual information* for training approach (MMI) (5.9) to three other training approaches, namely only using the provided base face images (base), using the base and all synthesized face images (synth.), and optimal synthesis selection guided by *maximum entropy* (ME) (5.8). We also compare our proposed optimal synthesis selection guided by *maximum mutual information* testing approach (testing: MMI), to cases when no synthesis is used at testing (testing: base).

Figure 5.3 shows visual results of optimal selections for both training and testing. Synthesis guided by ME tends to favor synthesis results at extreme poses and morphisms, whereas MMI selects a more balanced subset. In addition, MMI selects similar synthesis types at testing, and as we show below this leads to improved performance.

**Optimal synthesis selection.** Table 5.1 shows results when we vary the amount of synthesis types for training $k_{\text{train}}$. As expected, increasing the amount of synthesis at training leads to better performance (trends in rank-1 and rank-5 results for testing: base). In addition, incorporating synthesis at testing leads to greater increases in performance in all training setups. Furthermore, using our proposed MMI approach to select only 10 out of the 280 synthesis types at training, we achieved higher performance than using all 280 synthesis types.

As we observed in Table 5.1, highest performance is achieved when using synthesis guided by MMI during both training and at testing. Testing performance is sensitive to the amount of synthesis types $k_{\text{test}}$ at testing. We observed the opti-
Table 5.1: OFD rank-1 classification scores across testing setups for varying training setups and synthesis selection parameter for training, $k_{\text{train}}$. Results shown for calibration experiment with $N = 10$ training subjects and $k_{\text{test}} = 5$. Note that $k_{\text{train}} = 0$ is the same as training only with base images while $k_{\text{train}} = 280$ means training with all synthesized images. By incorporating MMI during training and testing, we achieve highest results while only using a small subset of the total synthesis data.

<table>
<thead>
<tr>
<th>$k_{\text{train}}$</th>
<th>ME</th>
<th>MMI</th>
<th>ME</th>
<th>MMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (base)</td>
<td>82.37</td>
<td>82.37</td>
<td>82.77</td>
<td>82.77</td>
</tr>
<tr>
<td>5</td>
<td>83.16</td>
<td>86.35</td>
<td>84.58</td>
<td>87.96</td>
</tr>
<tr>
<td>10</td>
<td>84.06</td>
<td>86.94</td>
<td>85.89</td>
<td>89.80</td>
</tr>
<tr>
<td>50</td>
<td>86.22</td>
<td>86.87</td>
<td>89.24</td>
<td>90.13</td>
</tr>
<tr>
<td>280 (synth.)</td>
<td>86.61</td>
<td>86.61</td>
<td>89.66</td>
<td>89.66</td>
</tr>
</tbody>
</table>

mal $k_{\text{test}}$ selection to be 5 and when we considered all possible synthesis types at testing, it slightly hindered performance (rank-1 classification scores were 89.80% and 86.38% for $k_{\text{test}} = 5$ and 35 respectively). We speculate this is possibly due to noise introduced from face synthesis, but considering computation costs and time at testing, lower values of $k_{\text{test}}$ are preferred.

**Effectiveness of synthesis.** Figure 5.5 fully demonstrates the effectiveness of using our proposed MMI approach during training and at testing. Performing any type of synthesis at training (Figure 5.5(b)) drastically improved rank-1 performance, where MMI and synth. showed the largest increase. Furthermore, the best results were achieved when MMI was also employed at testing (Figure 5.5(c)). Using MMI to select optimal synthesis types for training and testing, achieved nearly 93% rank-1 accuracy compared to 88% by using just the base images, and was able to efficiently use less than 5% of the total synthesis data, while achieving similar results to using all of the synthesis data (10 synthesis types vs. 280). This is further investigated in the Receiver Operating Characteristic (RoC) curve in Figure 5.6 and results are recorded in Table 5.2. Without touching or re-designing the state-of-the-art VGG-
Figure 5.6: RoC curves and performance for calibration experiments where $N = 10$, $k_{\text{train}} = 10$, and $k_{\text{test}} = 5$. (a) RoC curves for 4 training and testing approaches. Employed training and testing methods are represented in the legend, and are separated by a ‘-‘. Black solid line represents when only base images are used. The dotted lines represent cases when all synthesized images are used for training. The red line represents when synthesis was not performed at testing, while the blue is when MMI was employed at testing. Solid green line represents our proposed approach of using MMI both at training and at testing. The subplot shows a zoomed region of the RoC curves.

Face model, we were able to efficiently adapt it to the OFD dataset, drastically improving true positive rate (TRP) at 1% false alarm rate (FAR) from 0.79 to 0.88 and improve rank-1 classification from 80.70% to 89.19%.

5.5.2 Results on CASIA NIR-VIS 2.0

The CASIA NIR-VIS 2.0 dataset (Li et al. (2013a)) is the largest cross-spectrum face dataset available, containing 17,580 images across 725 subjects in both the near-infrared (NIR) and visual (VIS) spectrums. Benchmark results shown are rank-1 classification averages and standard deviations taken across from 10 validation sets where the NIR and VIS images are the probes and gallery respectively (Figure 5.7). The dataset contains images from two modalities (NIR and VIS) making this a very challenging dataset where VGG-Face achieves a 67.47% average rank-1 classification score. When enriching the training set, we treat each domain independently, in other
Table 5.2: Results on calibrating the VGG-Face model to the OFD dataset. MMI guided calibration for training and testing improved the rank-1 performance for VGG-Face from 80.70% to 89.19% without altering the DNN model.

<table>
<thead>
<tr>
<th>Training and testing setups</th>
<th>amount of training images</th>
<th><a href="mailto:TPR@0.01FAR">TPR@0.01FAR</a></th>
<th>rank-1 accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_{train}$</td>
<td>$k_{test}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 (base)</td>
<td>0 (base)</td>
<td>20</td>
<td>0.81</td>
</tr>
<tr>
<td>280 (synth.)</td>
<td>0 (base)</td>
<td>2,820</td>
<td>0.86</td>
</tr>
<tr>
<td>280 (synth.)</td>
<td>5 (MMI)</td>
<td>2,820</td>
<td>0.87</td>
</tr>
<tr>
<td>10 (MMI)</td>
<td>5 (MMI)</td>
<td>120</td>
<td>0.88</td>
</tr>
</tbody>
</table>

words, we do not perform synthesis on images from different domains. Specifically, for every subject we first randomly sample $M = 3$ images in each modality. Across the $M$ images we perform combinations of 6 Poisson face-swappings, 3 person-specific morphisms, and 3 pose variations. In total, we generate 324 different synthesis types for each subject. In the triplet embedding optimization, we also preserve domain information by requiring the anchor selection to be from a different domain than the positive and negative selections.

Figure 5.7 shows the top 4 synthesis types selected for testing. We record results in Table 5.3 and compare with multiple state-of-the-art results on the CASIA NIR-VIS 2.0 benchmark. Notably, we are able to adapt the VGG-Face and drastically improve rank-1 scores from 67.47% to 84.43% and outperforming many other works cited.

5.6 Conclusion

In this chapter, we proposed approaches for intelligent synthesis selection during training and testing. These approaches exploit face synthesis methods, allowing for more efficient training and improved testing performance. We demonstrated the effectiveness of the approaches through calibration experiments on the standard
Figure 5.7: Examples of images and optimal synthesis types from CASIA NIR-VIS 2.0 dataset. First column shows gallery (top) and probe (bottom) images from the dataset. The remaining columns are synthesis results from the provided base images chosen at testing by MMI.

Table 5.3: Results on CASIA NIR-VIS 2.0 benchmark. Lezama et al. (2017)$^2$ is in reference to the reported results of learning a low-rank embedding to the VGG-Face model.

<table>
<thead>
<tr>
<th></th>
<th>rank-1 accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lezama et al. (2017)$^1$</td>
<td>89.59 ± 0.89</td>
</tr>
<tr>
<td>Yi et al. (2015)</td>
<td>86.16 ± 0.98</td>
</tr>
<tr>
<td>Saxena and Verbeek (2016)</td>
<td>85.90 ± 0.90</td>
</tr>
<tr>
<td>Lu et al. (2015)</td>
<td>81.80 ± 2.30</td>
</tr>
<tr>
<td>Lezama et al. (2017)$^2$</td>
<td>80.69 ± 1.02</td>
</tr>
<tr>
<td>Juefei-Xu et al. (2015)</td>
<td>78.46 ± 1.67</td>
</tr>
<tr>
<td>Jin et al. (2015)</td>
<td>75.70 ± 2.5</td>
</tr>
<tr>
<td>VGG-Face</td>
<td>67.47 ± 1.73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training and testing setups</th>
<th>rank-1 accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_{train}$</td>
<td>$k_{test}$</td>
</tr>
<tr>
<td>0 (base)</td>
<td>0 (base)</td>
</tr>
<tr>
<td>72 (MMI)</td>
<td>0 (base)</td>
</tr>
<tr>
<td>324 (synth.)</td>
<td>0 (base)</td>
</tr>
<tr>
<td>324 (synth.)</td>
<td>5 (MMI)</td>
</tr>
<tr>
<td><strong>72 (MMI)</strong></td>
<td><strong>5 (MMI)</strong></td>
</tr>
</tbody>
</table>
and the challenging CASIA NIR-VIS 2.0 datasets. We outlined scenarios that
required a state-of-the-art DNN to be calibrated, and showed the impact of our
approaches both during training and at testing. For future work, we intend to shift
our optimal selection method to be integrated directly into the embedding algorithm,
allowing for a more active synthesis approach.
Conclusion

In this dissertation, we explored novel methods for behavioral analysis and focused on early risk marker identification for autism. We presented multiple contributions including a method for pose-invariant facial expression recognition, a self-contained mobile application for behavioral analysis, and a framework to calibrate a trained deep model with data synthesis and augmentation. We utilized multi-modal features at training and exploited cross-modal relationships during testing for pose-invariant facial expression recognition. We then extended our pose-invariant facial expression recognition method and presented other methods to characterize a multitude of risk behaviors related to risk marker identification for autism. To this end, we also developed a self-contained, closed-loop, mobile application that records a child’s face while he/she is watching specific, expertly-curated movie stimuli and automatically analyzes the behavioral responses of the child. We validated our methods to those of expert human raters, and using the developed and validated methods, we presented findings on group differences for behavioral risk markers for autism and interactions between motivational framing context, facial affect, and memory outcome. Lastly, we presented a framework to extend current state-of-the-art methods for face anal-
ysis. This framework optimally selects synthesis variations and employs synthesis both during training and at testing, leading to more efficient training and better performance.
Bibliography


109


111
Biography

Jordan Hashemi was born in Minnesota, in April of 1989. He spent most of his childhood in New Brighton, MN. He received his BEng from the Department of Biomedical Engineering and his MSc from the Department of Electrical Computer Engineering at the University of Minnesota in 2011 and 2013, respectively. He was awarded the Kristina M. Johnson Fellowship award in 2015. He was also part of a great inter-disciplinary team at Duke that was awarded the Blue Ribbon Team Award in 2015. His research interests include applied computer vision, machine learning, and behavioral coding analysis. He was one of the first graduate student in the Information Initiative at Duke (IID). He was also part of The Lives of Things project at Duke that created an interactive exhibit at the Nasher Museum of Art.

He expects to receive his Ph.D degree from the Department of Electrical and Computer Engineering at Duke University in the Spring of 2018.

His representative publications include (but not limited to):

