Are You Paying Attention? Classifying Attention in Pivotal Response Treatment Videos

Corey D C Heath, Hemanth Venkateswara, Sethuraman Panchanathan
Arizona State University
699 S Mill Ave, Tempe, AZ, USA, 85281

{corey.heath,hemanthv,panch}@asu.edu

Abstract

Pivotal response treatment (PRT) has been empirically shown to aid children with autism spectrum disorder ASD improve their communication skills. The child's primary caregivers can effectively implement PRT when provided with training and support, leading to greater opportunities for the child to improve. Utilization of computer vision technology is a critical component of creating more opportunities to support individuals implementing PRT. Automatically extracting data from videos of caregivers' interactions with their child during PRT sessions would alleviate the human effort required to provide assessment and feedback, which would allow experts to provide greater support to more individuals. Additionally, this data could be used to provide immediate automated feedback. The process of extracting data from PRT videos is complicated and provides excellent context for a computer vision challenge. PRT videos consist of 'in-the-wild' conditions of dyadic interactions recorded on ubiquitously available devices, and vary in filming quality.

The challenge presented tasks researchers with inferring the child's attention state in relation to the caregiver in the video based on body pose information and visual cues. Approaches will be evaluated based on accuracy metrics, however, the algorithm's speed is also important. Having fast algorithms will reduce the time between performance and assessment, allowing for greater opportunities to situate feedback in the context of the learning activity. Low-power solutions are also necessary to accommodate delivering results on mobile devices.

1. Introduction

Diagnosis rates for autism spectrum disorder (ASD) have shown a progressive increase in recent history [4]. Characteristically, children with ASD exhibit challenges in developing communication and social skills. Applied behavioral analysis (ABA) techniques, particularly pivotal response treatment (PRT), have been empirically shown to help children with ASD improve verbal communication skills [13, 16, 17, 14, 25, 19, 15, 18, 29]. PRT is a naturalist implementation of ABA methodology that involves an interventionist injecting learning opportunities into daily activities. In regard to developing child communication skills, this typically manifests as the interventionist identifying an object or activity the child is currently motivated by, and presenting learning opportunities based on this activity. This allows the interventionist to capitalize on the child's intrinsic motivation to continue the activity as a means of motivating compliance with the learning objectives.

Implementing PRT follows a guideline based on three phases of an interaction between an interventionist and a recipient. Often referred to as the ABC's of PRT, these phases consist of an antecedent, a behavior, and a consequence [20]. The antecedent represents the actions of the interventionist that create an opportunity for the recipient to exhibit a desired behavior. This involves the interventionist identifying the object or activity the recipient is motivated by, gaining the recipient's attention by exerting control over the activity, and presenting a learning opportunity in language at the recipient's level of understanding. The behavior is how the recipient responds to the antecedent. In a positive interaction, the behavior will be a sufficient attempt at performing the desired task, such as attempting to speak a prompted word. It is important to note that in communication skill development, a successful attempt is dependent on the skill level of the recipient. If the child has low verbal skills, an acceptable response may only consist of an individual phoneme from the word the interventionist prompted. The consequence is how the interventionist responds after the behavior exhibited by the recipient. For valid attempts, the interventionist should immediately continue the motivating activity and provide praise in order to

reinforce the correct behavior. If the behavior demonstrated by the recipient is insufficient, the interventionist should hold the motivating activity until a proper response is provided.

The advantages of training primary caregivers to conduct PRT with their child has been explored and shown to have positive effects in improving communication and social skills [21, 2, 9, 31, 5, 10, 26]. Utilizing caregivers as interventionists capitalizes on the greater amount of time the caregiver spends with their child, as opposed to relying on professional behavior analysts as the sole interventionist. In addition to improving child communication skills, caregivers also report lower stress and more positive affect for both them and their child after using PRT.

Training caregivers to implement PRT is problematic, particularly for residents of rural communities. PRT implementation training is primarily offered at autism resource centers, which may only be in urban centers. The training programs offered by these centers can be costly and time-intensive. These costs could make the training unobtainable for many individuals who would benefit from the services.

Exploring how to aid in the support and training of caregivers is a challenge that can be alleviated through the use of technology. In addition to the time investment required by caregivers to receive training, the clinicians who provide the training are also limited by these constraints. As diagnosis levels for ASD increase, clinical resources are stretched thin. PRT implementation training is based on analysis of both the interventionist and the recipient. Reviewing and providing adequate feedback to help the interventionist improve requires a large portion of the clinician's time. This makes providing support after training programs difficult. Current and future video processing technology can play a role in alleviating the manual costs involved in analyzing the videos and extracting important content for providing feedback. This could be used to reduce the time required by clinicians to review caregiver's PRT implementation. Additionally, data could be extracted from the videos to provide metrics and automated feedback that would aid the caregiver in improving fidelity to implementation.

Currently, caregiver fidelity to PRT implementation is assessed by clinicians through the evaluation of video probes depicting the caregiver using PRT with a child. These videos are scored in one minute increments based on categories that correspond to the antecedent, behavior, and consequence described above. The category of concern for this challenge is referred to as creating an 'opportunity to respond.' Identifying if an 'opportunity to respond' was provided requires analysis of the attention state of the child along with the language the interventionist uses in his or her instruction. This challenge tasks researchers with inferring the attention state of the child from the video. Creating this challenge would facilitate greater research into:

- Developing low power solutions to complex computer vision problems involving multiple human actors and 'in-the-wild' conditions
- Analyzing dyadic interactions
- Inferring human behavior from visual media
- Detecting states of attention and joint attention
- Providing progress toward creating automated tools will aid behavior analysts in supporting individuals working with child with ASD

2. Description of the Challenge

Creating a grand challenge would provide greater exposure to the problem of detecting and classifying dyadic interactions under 'in-the-wild' conditions. For the challenge, a dataset of PRT video would be created. The dataset would identify training and validation sets and task researchers with training classification models that provide the best performance. The performance needs to be assessed for accuracy and validation execution times. Low-power solutions are important for capitalizing on ubiquitous mobile technologies for processing data and returning results.

The goal of the challenges is to find accurate, efficient, and quick classification algorithms for detecting attention in videos under 'in-the-wild' conditions. The primary task in the challenge is to examine PRT video probes to identify when the child is attending the caregiver, inattentive to the caregiver, or if the child and caregiver are in a shared, or joint, attentive state. This information could then be used to automatically annotate videos, create video clips of important interactions, and provide metrics that could be used to provide performance-based feedback. Detecting attention involves inferring dyadic human behavior based on the individuals in the video frames, along with relevant objects. Ease-of-use is a concern for these types of systems. To maximize the number of users, the system needs to be based on ubiquitous technologies. Reflecting this, videos will represent 'in-the-wild' scenarios recorded on common technologies such as cell phones or hand-held cameras.

The data used in the challenge will be videos of individuals implementing PRT with children with ASD. The age range of participating children would likely be between two and 10 years old. A variety of environments and interactions would need to be depicted. Including this variety would address PRT's inherent dependency on child selected activities. These videos will be annotated for attention at one second intervals by behavior analysts or trained clinical professionals.

Evaluating solutions for the challenge needs to consider classification accuracy, system requirements, and processing times. In order for the solution to be beneficial as part

of a feedback system, a high level of accuracy is needed. Implementation system requirements for classification also need to be addressed to limit exclusionary constraints. The solution should run on consumer available technology, particularly mobile systems. Mobile systems are ideal for several reason. The ubiquity of mobile devices means that this will likely be the most convenient method for recording data. Having the classification system available on the device would eliminate the need for transferring data, making the process more secure. As the data includes videos of children, caregivers may be apprehensive about uploading it to a remote location. Using mobile devices also helps facilitate the naturalistic aspects of PRT. PRT does not have a structured implementation regarding activities and environments - it should be able to be implemented anywhere at any time. Mobile devices would give interventionists more freedom. As mobile technology is both the most likely data capture device, and is the equipment the interventionist is likely to carry with them, providing results on the device would allow for more immediate presentation of feedback. Timely assessment is an important aspect of providing feedback. Ideal solutions should present results quickly to allow performance metrics to be reviewed while the learning context is in recent memory. Real-time classification could also be utilized to provide feedback that would allow an individual to immediately correct his or her behavior. Evaluation of execution time will focus solely on validation processing. Offline training is permitted, and its execution time is not a primary concern for the challenge.

Mobile devices should be incorporated into the evaluation of the challenge entries. The implementation should be flexible to accommodate different approaches. Classification models may be trained on any system, however, validation should be undertaken on a mobile device. This could consist of processing a video stream as it is recorded, or processing a video file post-recording.

3. Beneficiaries of the Technology

Clinicians and caregivers that would like to learn PRT are the direct beneficiaries of exploring the technologies in this challenge. Clinicians would benefit from having access to tools that would reduce the time they need to invest to analyze performance and provide feedback to individuals learning PRT [11]. This would give the clinician the opportunity to support more clients, and provide more in-depth assessments.

Caregivers would benefit from having more resources available for learning PRT and receiving ongoing support that would aid them in adapting the process to new skills. The use of technology would also enable more opportunities for remote training and support for rural communities. In addition to clinicians and caregivers, autism and ABA researchers would benefit from the technology by having an additional tool for assessments and metric-gathering.

Beyond PRT, dyadic attention detection could be useful in other educational scenarios, such as classrooms, tutor sessions, or athletics training and coaching. The technology could also be useful in business environments, particularly in automating evaluation of job interviews.

The computer vision research community also benefits from the creation of a difficult dataset and challenging task requirements. The task requires researchers to make inferences regarding dyadic human behavior based on visual cues. This is a complex problem that may incorporate many sub-problems including human pose detection and body segmentation, gaze estimation, facial expression recognition, individual and dyadic activity detection, and engagement and attention detection. Additionally, due to the dataset, algorithms for approaching these sub-problems need to be robust against occlusion, low-resolution, incomplete data, and low training set sample representation.

4. Challenges in Processing PRT Video Probes

The challenges in processing the PRT videos are largely due to the nature of PRT. PRT is dependent on the recipient selecting the primary activity for the session. This means that the activity is not known when planning automated analysis strategies. Strategies for detecting attention need to be able to generalize the signs of attention and extract them from the interaction in order to make a successful classification. Having participants with ASD as primary subjects compounds the issue of extracting attention. Children with ASD do not exhibit overtly visible signs of attention. Additionally, play activities will be commonly depicted which may consist of quick, erratic movements.

The need for the system to accommodate ubiquitous recording technologies presents additional challenges. The most common expected scenario for the videos is the caregiver interacting with his or her child in a home environment. This means the recordings will exhibit a range of qualities based on the recording equipment, the skill of the camera operator, and the visual-noise present in the environment. Issues such as occlusion and unstable camera should be anticipated. Additionally, a procedure for excluding individuals not involved in the PRT interaction needs to be considered. The two-dimensionality of the video also poses a problem, particularly for inferring attention based on visual estimations.

Despite these challenges, the task is accomplishable. Currently, clinicians evaluate PRT videos and extract information based on the child's attention. This is demonstrated by the use of PRT fidelity score sheets both in research studies and in training practice at resource centers.

Handling data may also be a concern for some users as the videos contain children. The ability to extract key information from the video on the recording device would

prevent the video from needing to be distributed to a separate system for processing. This provides users with more options on managing access to the videos.

5. Description of the Data

The videos will be recorded using technology that is accessible and requires little preparation before usage. This is meant to replicate a final implementation environment that emphasizes accessibility for users. This will result in untrimmed 'in-the-wild' data. Ideally, the videos will depict a caregiver-child dyad participating in play activities.

A dataset for PRT video probes has been explored by [12]. This dataset consisted of 14 videos, each approximately 10 minutes in length. The videos depicted baseline and post-treatment sessions for seven parent-child dyads.

Thirty frame segments, representing one second of video, were extracted and labeled based on the child's attention state. The labels included when the child was attentive, inattentive, or the parent and child are engaged in a shared activity. These states are defined based on the control of the motivational activity. The child is considered inattentive if he had sole control of the motivational activities, was engaged in play activities, or ambulatory. Additionally, the child would be considered inattentive if he was exhibiting tantrum behavior, or was engaged in self-stimulation. The child was considered to be attentive when his focus was aimed at the parent. This generally means the parent had control of the motivational activity. Signs the child was attentive to the parent include: visual focus on the adult, particularly on the adult's face or an object held the adult's hands; body is oriented toward the adult; and, not being engaged in movement or other activities [28]. The final state, shared attention, represented the parent and child being engaged in a dyadic activity. This was distinguished from the attentive state by the level of disruption that is needed for the parent to exert control over the activity to present a learning opportunity. In the attentive state, the parent seizing control of the activity causes the child to cease the activity and change his focus toward the parent.

These states of attention can be identified in a PRT training video [30]. The attentive state (figure 1) is demonstrated when the interventionist, the woman on the left, presents a motivating object to the recipient, the woman on the right. The recipient is visually focused on the interventionist, particularly the puppet in the interventionist's hand. Conversely, after the recipient gains control of the object, she enters the inattentive state (figure 2). The recipient is now engaged with the object and is less likely to respond to instructions from the interventionist. The shared attention state (figure 3) in this scenario is similar to the attentive state. The interventionist now has her own puppet and is engaging in the play activity with the recipient. This allows the interventionist the ability to present learning opportunities with less disruption to the play activity.

Figure 1. Screenshot from a PRT training video. The recipient (right) is in an attentive state, as indicated by her looking directly at the interventionist (left).

Figure 2. Screenshot from a PRT training video. The recipient (right) is engaged with the puppet and would be less receptive to instructions from the interventionist (left), indicating an inattentive state.

The dataset illustrates many of the inherent problems this challenge tasks researchers to consider. A wide array of activities are explored throughout the videos, including playing with toy cars, watching videos on a cell phone, and spinning in an office chair. Each of these activities exhibits different configurations of the parent-child dyad that relate to different signs of attention from the child. The imbalanced nature of the domain space is problematic for training classification models. Table 1 shows the number of 30-frame segments from the video that were ascribed to each class. Segments were ignored when either the adult or child were entirely out of the camera view. This table illustrates that the majority of samples are in the Inattentive class, however, this is largely dependent on the activity depicted in the video. The Dyad2 Post and Dyad3 Base videos are outliers, consisting primarily of Shared attention segments. This is because the primary activities being participated in

Figure 3. Screenshot from a PRT training video. The interventionist (left) and recipient (right) are engaged in a joint activity, playing with puppets. This is a demonstration of a shared attentive state.

are watching a video, and playing a game, respectively, and spanned the entirety of the video.Additionally, the dataset illustrates the periods of occlusion, superfluous individuals, and unstable camera movements.

Table 1. The attention class label counts for each video probe [12].

Video Probe	Attentive	Shared	Inattentive	Ignored
Dyad1 Base	182	43	371	5
Dyad1 Post	170	23	266	156
Dyad2 Base	178	4	254	170
Dyad2 Post	11	585	14	0
Dyad3 Base	146	258	190	10
Dyad3 Post	203	101	133	167
Dyad4 Base	80	0	278	260
Dyad4 Post	261	22	285	33
Dyad5 Base	35	144	415	17
Dyad5 Post	144	66	372	29
Dyad6 Base	95	180	215	125
Dyad6 Post	135	26	317	127
Dyad7 Base	94	110	167	236
Dyad7 Post	119	246	221	24

The effect of these challenges was examined in [12] using OpenPose [3] to identify individuals in the videos and provide locations of body and facial features. Table 2 shows the proportion of feature points identified in each frame, along with the report confidence score. On average only 66% of the body points were recognized in a given frame. For these points, the average confidence level reported by OpenPose was 56%. Facial features were more problematic. While OpenPose detected facial features 70% of the identified individuals in the frames, the average confidence for these points was only 23%. The expectation is that the videos will depict dyadic interactions, but the evaluation of the PRT video probes shows that in the majority of frames only one individual is identified (figure 4A). Using post-processing procedures, missing data can be approximated to favor having two individuals present. This is accomplished by constructing missing points based on known data from previous and future frames. When more than two people are detected in the frame, information from previous frames, along with the proximity to other individuals, and recognition confidence, are used to exclude persons not believed to be involved in the interaction. Figure 4B shows this methodology has successfully promoted having only two individuals in each frame.

Table 2. The proportions and confidence levels for body and facial point detection from OpenPose for each video probe [12].

Video Probe	Body Det.	Body Conf.	Face Det.	Face Conf.
Dyad1 Base	0.72	0.58	0.76	0.13
Dyad1 Post	0.62	0.55	0.6	0.08
Dyad2 Base	0.62	0.56	0.87	0.36
Dyad2 Post	0.69	0.59	0.8	0.35
Dyad3 Base	0.63	0.5	0.85	0.26
Dyad3 Post	0.53	0.56	0.79	0.33
Dyad4 Base	0.74	0.57	0.79	0.23
Dyad4 Post	0.74	0.56	0.83	0.23
Dyad5 Base	0.72	0.57	0.95	0.34
Dyad5 Post	0.72	0.53	0.78	0.17
Dyad6 Base	0.55	0.54	0.63	0.1
Dyad6 Post	0.59	0.52	0.63	0.13
Dyad7 Base	0.62	0.55	0.82	0.24
Dyad7 Post	0.68	0.59	0.91	0.32

Although this dataset provides insight into the problem, more data is needed to create a stable solution. Including more activities would aid in the creation of classification models that are more robust. Demographically, the dataset solely consisted of mother-son dyads, with the child being between the ages of two and five years old. More diversity in demographics, age, and exhibited verbal and social skills is necessary.

The distribution agreement for the dataset is a major limitation. The dataset was collected under a limited research agreement and is not available for distribution outside of the research and resource center where it was collected.

6. Methods to Acquire and Annotate Data

Collecting additional data would likely rely on the collaboration with psychology and behavior analysis departments at universities as well as community autism research and resource centers. Videos would be collected from caregivers and children participating in research studies involving PRT or attending PRT training programs. Compen-

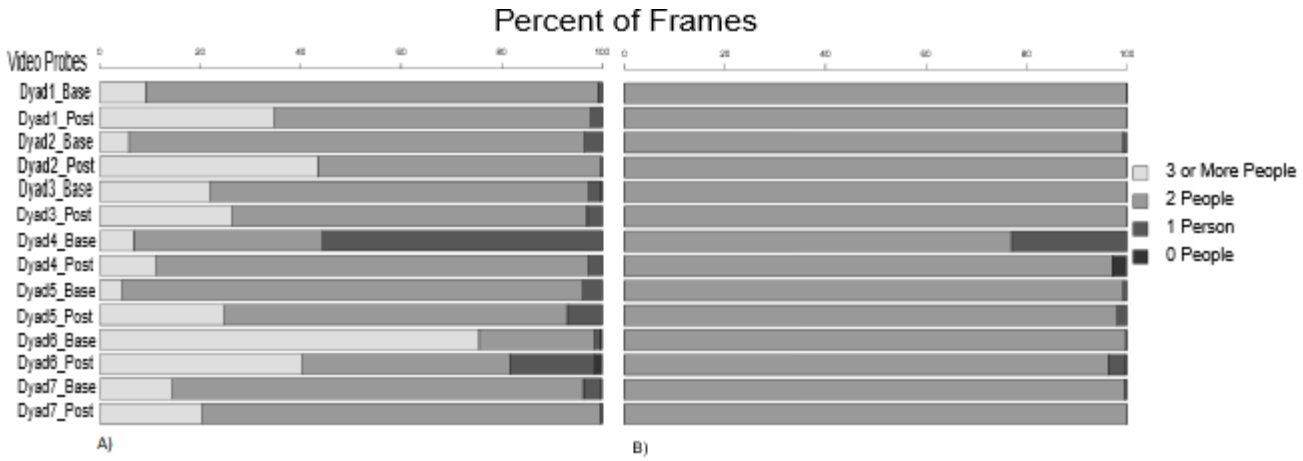

Figure 4. A) The bar graph shows the number of people detected by OpenPose in each video. The bars illustrate the percent of frames by the number of people detected. B) Shows the percentages after processing [12].

sation could be provided to participants to cover training course enrollment costs and other related fees. The authors of this challenge are actively seeking grant funding to build this dataset.

Data capture for the new dataset will follow procedures currently used in PRT research and training. Videos will consist of 10-minute PRT sessions between the caregiver and the child. These videos will be filmed from an exocentric perspective, most likely by a third party. The capturing device would be a hand-held camera or cell phone. Participants would be give the option of using their own device or one provided by the researchers. A minimum of three videos per participating dyad would need to be collected. These would consist of a pre-training baseline video, immediate post-training video, and an additional post-training video two weeks after the training has concluded.

Labeling the datasets should be undertaken by at least two behavior analysts. As behavior can be subjective, having more professionals involved in the labeling process will aid by providing a means for comparing and corroborating labels. This will illustrate explicit moments or attention versus more enigmatic moments. In addition to attention labels for video segments, the analysts should provide traditional PRT fidelity metrics.

7. Description of Copyright and Privacy

The dataset includes children and must be manage in a responsible manner. Collecting the data should be undertaken with a clear end user licence agreement (EULA) stating how the data will be used, managed and distributed. The recommendation would be to require researchers to apply for access to the data and have their project approved by an institutional review board.

8. Methods for Scoring Solutions

The challenge represents a classification task. Scoring approaches to solving the challenge would involve utilizing accuracy, precision, and recall metrics. These would be calculated by comparing the proposed algorithms predictions against labeled samples from a test or validation set. This could occur at several levels of granularity. [12] focused on classifying one-second video segments, however, they also identified that this level of detail may not be necessary for achieving adequate solutions to the problem. The current human-created evaluations for fidelity are based on one-minute video segments. Valid solutions could be scored against human collected fidelity metrics to determine if the automated algorithm could recreate the human assessments. Additionally, it was noted in [11] that clinicians felt that for feedback purposes, 10 - 15 second segments is the smallest useful increment. Examining each of these levels of precision allows for flexibility in the approaches that could be implemented.

In addition to accuracy metrics, the goal of the challenge is to move toward creating a feedback model. For the feedback to be successful, it needs to be presented as close to when the behavior occurred as possible in order to situate the feedback in context. This allows the learner to easily recall their actions in the situation and map the assessment to the behavior, which reinforces positive behaviors and provides insight into erroneous actions. The approach of reviewing video immediately after PRT sessions was explored by [24, 27], and shown to be an effective method for promoting learning and instilling self-efficacy. In terms of the challenge, this means the speed of classification algorithms is also an important consideration. The goal should be to create solutions that can present results soon after the video recording is finished. Performing classification near to real-

time would also be important and provide the opportunity to instruct the learner on his or her behavior, allowing the learner to take immediate corrective actions.

As the solution is intended for mobile platforms, energy efficiency is an important metric for evaluating and comparing approaches. Energy consumption will be reported in Watts per hour.

9. Constraints on Hardware

This hardware constraint is one of the particular challenges that prevents a satisfactory solution being obtainable with currently technology, however, it is reasonable to assume that accommodating technology will be available in the next five years. In [12] the proposed solution for attention detection utilized OpenPose [3] to identify individuals and extract body pose information. OpenPose currently has high hardware demands for producing results quickly. The challenge seeks to follow an implementation format based on ubiquitous technologies. The recording equipment used for creating the dataset should be limited to cell phone or hand-held cameras. Hardware used for data processing and classification should follow similar constraints. Hardware systems used in the challenge should emphasize interfacing with consumer devices, particularly mobile products.

Mobile processing power is likely to increase greatly in the next five years. For this challenge, the increased computing power will provide more opportunities to implement classification algorithms directly on the recording device. There are several benefits to a local solution. It would make the application easier to use and develop by not needing to export data to a remote location for processing. Additionally data would be more secure, as it would not need to leave the user's device. Privacy may be an important concern for some users, especially because the data involves children.

Along with improvements in processing power over the next five years, the expectation is that video processing research will continue to find new, more efficient methods for data extraction and analysis. This could include more powerful deep learning recognition models, greater usage of transfer learning models, and more data related to this problem.

10. Maximum and Minimum Requirements

There are no current maximum or minimum requirements for the challenge. The current baseline accuracy for similar task was presented in [12]. This was 44% on a three-class classification task. No baseline metrics are available for validation execution time or energy consumption.

11. Differences Between Current Competitions

The presented challenge is congruent with tasks presented in previous low-power image recognition challenges (LPIRC). As with previous challenges, favorable results requires systems to analyze an image and detect important objects in a limited amount of time on a constrained system. The primary difference between this challenge and current competitions is the emphasis on classifying human behavior. This encompasses recognition of people, their pose information, and how that information is choreographed in a scene along with other static and dynamic objects. This builds upon previous LPIRC that focus on object recognition, classification, and localization [23, 6, 8, 1, 22]. The proposed challenge tasks systems using image recognition to make inferences about the attention state of human participants. The EmotiW competition is similar to the proposed challenge [7]. EmotiW focuses on 'in-the-wild' emotion and engagement detection, however, EmotiW examines single adult individuals, fixed camera perspective focusing on the individual's face and upper torso, and multimodal data. Conversely, the PRT video challenge examines dyadic behavior, has an exocentric camera perspective encapsulating a dynamic scene, and relies on visual data to detect attention.

12. Conclusion

Detecting attention in video data is a difficult task. Providing constraints on the data recording devices and data processing systems adds additional challenges, however, emphasizing these limitations leads to approaches that will be easier to adapt to mainstream implementation. This challenge focuses on the application of video classification to a distinct usage - evaluating PRT video probes. Approaching this task provides an opportunity to benefit the community by providing resources that aid individuals in learning ABA methodologies that have been proven to foster the development of communication skills in children with ASD. Additionally, creating automated data extraction and feedback tools will allow for greater support of individuals who do not have adequate access to support resources.

13. Acknowledgments

The authors thank Arizona State University and the National Science Foundation for their funding support. This material is partially based upon work supported by the National Science Foundation under Grant No. 1069125 and 1828010.

References

[1] Sergei Alyamkin, Matthew Ardi, Achille Brighton, Alexander C Berg, Yiran Chen, Hsin-Pai Cheng, Bo Chen, Zichen Fan, Chen Feng, Bo Fu, and others. 2018 low-power image recognition challenge. *arXiv preprint arXiv:1810.01732*, 2018.

[2] Mary J Baker-Ericzn, Aubyn C Stahmer, and Amelia Burns. Child demographics associated with outcomes in a community-based pivotal response training program. *Journal of positive behavior interventions*, 9(1):52–60, 2007.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[4] CDC. Data and statistics on autism spectrum disorder | CDC. `https://www.cdc.gov/ncbddd/autism/data.html`, 2019.

[5] Jamesie Coolican, Isabel M Smith, and Susan E Bryson. Brief parent training in pivotal response treatment for preschoolers with autism. *Journal of Child Psychology and Psychiatry*, 51(12):1321–1330, 2010.

[6] Erik Debenedictis, Yung-Hsiang Lu, Alan Kadin, Alexander Berg, Thomas Conte, Rachit Garg, Ganesh Gingade, Bichlien Hoang, Yongzhen Huang, Boxun Li, and others. Rebooting computing and low-power image recognition challenge., 2016.

[7] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 653–656. ACM, 2018.

[8] Kent Gauen, Rohit Rangan, Anup Mohan, Yung-Hsiang Lu, Wei Liu, and Alexander C Berg. Low-power image recognition challenge. In *2017 22nd Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 99–104. IEEE, 2017.

[9] Jill N Gillett and Linda A LeBlanc. Parent-implemented natural language paradigm to increase language and play in children with autism. *Research in Autism Spectrum Disorders*, 1(3):247–255, 2007.

[10] Antonio Y Hardan, Grace W Gengoux, Kari L Berquist, Robin A Libove, Christina M Ardel, Jennifer Phillips, Thomas W Frazier, and Mendy B Minjarez. A randomized controlled trial of pivotal response treatment group for parents of children with autism. *Journal of Child Psychology and Psychiatry*, 56(8):884–892, 2015.

[11] Corey DC Heath, Tracey Heath, Troy McDaniel, Hemanth Venkateswara, and Sethuraman Panchanathan. Designing a user interface for multimodal analytics extracted from videos of parent implemented pivotal response treatment sessions. In *ACM SIGACCESS Conference on Computers and Accessibility*, 2019.

[12] Corey DC Heath, Hemanth Venkateswara, Troy McDaniel, and Sethuraman Panchanathan. Detecting attention in pivotal response treatment video probes. In *International Conference on Smart Multimedia*, 2018.

[13] Lynn Kern Koegel, Stephen M Camarata, Marta Valdez-Menchaca, and Robert L Koegel. Setting generalization of question-asking by children with autism. *American Journal on Mental Retardation*, 102(4):346–357, 1997.

[14] Lynn Kern Koegel, Cynthia M Carter, and Robert L Koegel. Teaching children with autism self-initiations as a pivotal response. *Topics in language disorders*, 23(2):134–145, 2003.

[15] Lynn Kern Koegel, Robert L Koegel, Israel Green-Hopkins, and Cynthia Carter Barnes. Brief report: Question-asking and collateral language acquisition in children with autism. *Journal of Autism and Developmental Disorders*, 40(4):509–515, 2010.

[16] Lynn Kern Koegel, Robert L Koegel, Joshua K Harrower, and Cynthia Marie Carter. Pivotal response intervention i: Overview of approach. *Journal of the Association for Persons with Severe Handicaps*, 24(3):174–185, 1999.

[17] Lynn Kern Koegel, Robert L Koegel, Yifat Shoshan, and Erin McNerney. Pivotal response intervention II: Preliminary long-term outcome data. *Journal of the Association for Persons with Severe Handicaps*, 24(3):186–198, 1999.

[18] Robert L Koegel, Jessica L Bradshaw, Kristen Ashbaugh, and Lynn Kern Koegel. Improving question-asking initiations in young children with autism using pivotal response treatment. *Journal of autism and developmental disorders*, 44(4):816–827, 2014.

[19] Robert L Koegel, Ty W Vernon, and Lynn K Koegel. Improving social initiations in young children with autism using reinforcers with embedded social interactions. *Journal of autism and developmental disorders*, 39(9):1240–1251, 2009.

[20] Rebecca Landa. Early communication development and intervention for children with autism. *Mental retardation and developmental disabilities research reviews*, 13(1):16–25, 2007.

[21] Karen E Laski, Marjorie H Charlop, and Laura Schreibman. Training parents to use the natural language paradigm to increase their autistic children's speech. *Journal of Applied Behavior Analysis*, 21(4):391–400, 1988.

[22] Yung-Hsiang Lu. Low-power image recognition. *Nature Machine Intelligence*, 1(4):199, 2019.

[23] Yung-Hsiang Lu, Alan M Kadin, Alexander C Berg, Thomas M Conte, Erik P DeBenedictis, Rachit Garg, Ganesh Gingade, Bichlien Hoang, Yongzhen Huang, Boxun Li, and others. Rebooting computing and low-power image recognition challenge. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 927–932. IEEE, 2015.

[24] Suzanne Elaine Robinson. Teaching paraprofessionals of students with autism to implement pivotal response treatment in inclusive school settings using a brief video feedback training package. *Focus on Autism and Other Developmental Disabilities*, 26(2):105–118, 2011.

[25] Michelle R Sherer and Laura Schreibman. Individual behavioral profiles and predictors of treatment effectiveness for children with autism. *Journal of consulting and clinical psychology*, 73(3):525, 2005.

[26] Isabel M Smith, Helen E Flanagan, Nancy Garon, and Susan E Bryson. Effectiveness of community-based early intervention based on pivotal response treatment. *Journal of Autism and Developmental Disorders*, 45(6):1858–1872, 2015.

[27] Jessica Suhrheinrich and Janice Chan. Exploring the effect of immediate video feedback on coaching. *Journal of Special Education Technology*, 32(1):47–53, 2017.

[28] Jessica Suhrheinrich, Sarah Reed, Laura Schreibman, and Cynthia Bolduc. *Classroom pivotal response teaching for children with autism*. Guilford Press, 2011.

[29] Pamela Ventola, Hannah E Friedman, Laura C Anderson, Julie M Wolf, Devon Oosting, Jennifer Foss-Feig, Nicole McDonald, Fred Volkmar, and Kevin A Pelphrey. Improvements in social and adaptive functioning following short-duration PRT program: a clinical replication. *Journal of autism and developmental disorders*, 44(11):2862–2870, 2014.

[30] Virgir05. Pivotal response treatment PRT example. `https://www.youtube.com/watch?v=vZOS-aYRVOI`, 2015.

[31] Laurie A Vismara and Gregory L Lyons. Using perseverative interests to elicit joint attention behaviors in young children with autism: Theoretical and clinical implications for understanding motivation. *Journal of Positive Behavior Interventions*, 9(4):214–228, 2007.