Guess What?

Towards Understanding Autism from Structured Video Using Facial Affect

Haik Kalantarian¹ · Peter Washington¹ · Jessey Schwartz¹ · Jena Daniels¹ · Nick Haber¹ · Dennis P. Wall¹,²

Received: 15 April 2018 / Revised: 31 August 2018 / Accepted: 8 September 2018 / © Springer Nature Switzerland AG 2018

Abstract

Autism Spectrum Disorder (ASD) is a condition affecting an estimated 1 in 59 children in the United States. Due to delays in diagnosis and imbalances in coverage, it is necessary to develop new methods of care delivery that can appropriately empower children and caregivers by capitalizing on mobile tools and wearable devices for use outside of clinical settings. In this paper, we present a mobile charades-style game, Guess What?, used for the acquisition of structured video from children with ASD for behavioral disease research. We then apply face tracking and emotion recognition algorithms to videos acquired through Guess What? game play. By analyzing facial affect in response to various prompts, we demonstrate that engagement and facial affect can be quantified and measured using real-time image processing algorithms: an important first-step for future therapies, at-home screenings, and outcome measures based on home video. Our study of eight subjects demonstrates the efficacy of this system for deriving highly emotive structured video from children with ASD through an engaging gamified mobile platform, while revealing the most efficacious prompts and categories for producing diverse emotion in participants.

Keywords  Autism · Emotion · Mobile

1 Introduction

Autism spectrum disorder (ASD) is a developmental disability that affects 70 million children worldwide, including 750 thousand American children under the age of five.
The prevalence of ASD has increased in recent years, with the rate of incidence in American children now estimated to be 1 in 59 [32]. As ASD is considered a spectrum condition, symptoms vary from one child to another. However, most children with ASD struggle to make eye contact, recognize facial expressions, and engage in social interactions with their family and peers [11, 30, 38]. Other behaviors associated with ASD include maintaining stereotyped behaviors, poor motor skills, and difficulty with language [32]. In addition to the substantial burden that this condition places on children and families, it has been estimated that the annual costs associated with ASD reach almost $90 billion in the USA as a whole [32].

Standard methods of ASD diagnosis include the Autism Diagnostic Observation Schedule (ADOS) [25] and Autism Diagnosis Instrument Revised (ADI-R) [26]. These instruments are almost always administered in clinical facilities and require hours of administration by specialists—factors that contribute to delays in diagnosis and imbalances in coverage. This is particularly the case for families of lower socioeconomic status, where the average time between initial warning and diagnosis can exceed 18 months [3, 37]. By the age of eight, it is estimated that 27% of children remain undiagnosed, which is often beyond the timeframe in which therapy is most effective [7–10]. Possible alternative solutions that can ameliorate some of these challenges could be derived from digital and mobile tools.

In this paper, we present Guess What?: a mobile game we have developed and published on the Google Play Store [23] based on the mechanics of the popular charades game, Heads Up. In Guess What?, children act out various images shown on the phone while being recorded by their care provider, who attempt to guess the prompt. The parent labels the video by tilting the phone after they have correctly identified what their child is expressing, at which time another prompt is presented until time is up. Using this method of crowdsourced at-home video acquisition, we are developing a database of children with ASD as well as neurotypical children as they express themselves in response to various prompts: emojis, facial expressions, and more abstract prompts such as sports games, occupations, and animals.

Prior works [6, 35, 36] have demonstrated the efficacy of mobile video phenotyping approaches for children with ASD. Similarly, the game mechanics of Guess What? are well-suited for ASD research because they facilitate video capture in a semi-structured environment in which the child’s face is almost always centered correctly in the frame. Moreover, the primary deficits with which children with ASD struggle include emotion recognition, facial affect, and expressive language communication. These deficits can not only be detected but also measured using the format of a charades-style game as a method of understanding the impact of ASD on emotion recognition and representation [32]. Therefore, we predict that a comprehensive database of video data from a mobile game can become an invaluable resource for future computer vision-based ASD research such as therapeutic interventions, screening techniques, and the development of novel outcome measures.
To validate our platform and demonstrate the utility of using home video characterizing facial affect within the constraints of our mobile game, we have conducted an in-lab study in which eight children with ASD played Guess What?. Using facial recognition technology applied to video data, we analyzed the children’s performance and behavior as they respond to prompts in a variety of different categories: real faces, emoji, sports, occupations, and animals. The system architecture is shown in Fig. 1; our system collects videos as children react to various prompts shown on the phone while being recorded using the device’s front camera. The footage is uploaded to an Amazon AWS bucket, where it is stored along with metadata about the session and participant. In a post-processing step, various facial-expression recognition algorithms [19] are employed to analyze video frames individually, recognizing emotion and monitoring facial engagement in real-time during the game session.

The primary contributions of this paper are as follows:

– We describe our mobile game for crowdsourced video acquisition, Guess What?, based on the mechanics of the popular charades game, Heads Up [4]. This system is used to aggregate structured video from children in an interactive environment. Based on our prior studies on computer vision-based therapies using Google Glass [36], we believe this database of video footage can be used as a platform for future computer vision-based ASD research such as therapeutic interventions, at-home autism screening, and the development of novel outcome measures.

– We present an in-lab study in which children with ASD played Guess What?. Through this study, we explored the feasibility of this platform to collect emotion-rich social video from children between the ages of four and eleven.

– We apply face tracker and emotion recognition algorithms to acquired video from our in-lab study to explore several questions about how children responded to various elements of the game. These questions included:
  
  – Which categories elicit the most emotive and diverse facial expressions in children?
  – Which categories and prompts are the most challenging, and which are the easiest?
  – Are emoji representations of exaggerated emotions more suitable than photos of faces, for eliciting emotive facial expressions?
  – Which category is associated with the highest level of facial engagement?

This paper is organized as follows. Section 2 provides related work, primarily in applications of computer vision to ASD. Section 3 describes the system used to collect, annotate, and securely store video data on Amazon AWS servers. Section 4 describes an overview of algorithms used for emotion classification and meta-analysis of videos to screen for ASD. Section 5 describes our experimental methods, followed by results and discussion in Section 6. We describe study limitations in Section 7 and provide concluding remarks in Section 8.
Fig. 1 The architecture of the proposed system is shown above. Video is captured from a phone’s camera during a mobile charades game session and stored in a secure AWS server along with demographic information about the participant. Using histogram of oriented-gradients, frames from the video are classified according to the emotion displayed by the child at any given time throughout the session. Information about the child’s performance, as well as supplementary behavioral information automatically extracted from the video, are explored for their efficacy in evoking emotive facial expressions in participants.
2 Related Work

A large body of work in recent years has explored the use of technology to enhance therapy for children with autism spectrum disorder. For example, one study revealed that watching a television series designed to enhance emotion comprehension every day can improve the emotion comprehension of young children with ASD [18]. It is therefore plausible that regular exposure of a child to challenges that develop social acuity can have a positive long-term impact on their well-being. However, the most important takeaway of these high-tech aids and therapies within the context of our proposed system is the conclusion that children with ASD show high engagement with interactive gamified systems; these approaches are generally well received and present an exciting opportunity to make a positive impact on children with developmental delays.

One example of a novel therapy is the ECHOES project, which presents a technology-enhanced learning (TEL) environment that facilitates acquisition and exploration of social skills by typically developing children and children with ASD [2]. Multitouch interfaces for behavioral therapy have also been found to increase collaboration skills in children with ASD [1, 16, 17]. Humanoid robots used as catalysts for social behavior have also been shown to promote positive child-human interactions [15]. Frameworks for designing games to teach emotion to children with ASD have also been shown to be successful [28, 35, 36]. Robot therapy for ASD has been explored as one of the first applications for socially assistive robotics; Scassellati et al. found that social robots may be developed into useful tools for social skills and communication therapies, specifically by embedding social interaction into intrinsic reinforcers and motivators for children with ASD [31].

Interactive visual displays have also been developed for children with ASD. vSked, an interactive and collaborative visual scheduling system designed for children with ASD in elementary school classrooms, was found to reduce staff effort for using the visual cues and resulted in improvements in the perceived quality and quantity of communication and social interactions in the classroom [21, 40]. Techniques that involve verbal and visual prompting annotated on top of the physical objects used during therapies have also been explored in Mobis—a mobile augmented reality application that lets teachers superimpose digital content on top of physical objects [14]. MOSOCO, another mobile assistive technology, uses the visual supports of a validated curriculum to help children with ASD practice social skills in real-life situations [13]. Hayes et al. have determined design requirements for assistive technologies that use interactive visual supports to engage students and support parents simultaneously: flexibility, communication and collaboration capabilities for both children and parents, and parent support for programming and documentation of use [20].

There are also various augmented reality systems for children with ASD. The Superpower Glass Project is a Google Glass application which displays visual cues to a child with ASD about real-time facial expression data from a conversational partner [6, 35, 36]. Others have explored the use of wearable systems and affective computing as companion tools for social-emotional learning and the use of the recorded videos for defining a process to collect, segment, label, and use video clips from everyday conversations [29, 34].
Our system differs from many previous technological solutions for ASD behavioral support due to its focus on collecting data which could benefit a future population while providing immediate behavioral support. To support this goal, we gamify the data collection process, leveraging the gameplay mechanics of the popular mobile app *Heads Up*, which has been downloaded millions of times on both the Android and iOS platforms and has been well received by a variety of age groups [22]. Moreover, our system emphasizes challenging participants’ abilities of imitation and figurative interpretation of images in a social setting that includes both parent and child.

### 3 System Description

In this section, we describe *Guess What*—an Android mobile application for crowdsourcing the acquisition of egocentric video. This app is a medium through which

---

*Fig. 2* After providing consent, parents are asked to complete an optional survey providing the child’s first name, age, gender, race/ethnicity, and diagnosis. This data is retained on a secure Amazon DynamoDB database.
we are developing a database of video data that will be used to further our understanding of which prompts and images most effectively elicit emotion and reveal differences in facial affect between children with ASD and those with no prior diagnosis. These results will help us develop a better understanding of how this disorder affects children’s ability to interpret and imitate.

3.1 Game Mechanics

When users first download the app from the Google Play Store, they are instructed to create a new account at which time they are presented with three IRB-approved consent forms through which they indicate their willingness to share videos with researchers. If the primary care provider consents to share this information, they are directed to the form shown in Fig. 2 at which time they can optionally provide the child’s first name, age, gender, prior diagnosis, and country of origin. It is expected that the caregiver will answer these questions on behalf of the child.

Fig. 3  Players can select between several decks of prompts, each posing unique challenges with respect to difficulty of interpretation and expression. Options range from simple emotions and photos to abstract challenges such as sports games and occupations.
After account creation, parents are directed to the deck-selection screen shown in Fig. 3. This interface allows parents to decide which prompts will be shown to the child, which vary greatly in difficulty in order to appeal to a range of age groups and levels of developmental delay. The possibilities include emoji, animals, faces, sports, occupations, and a random category which selects among all of these. Naturally, the choice of prompts dramatically affects the quality of the incoming video data; showing the child photos of real people is asking them to imitate the expression within the photo, while showing a more abstract image such as that of an astronaut challenges their creativity and interpretation skills. Each deck provides unique information: we can directly apply emotion recognition algorithms to video footage of a child being shown an emotive prompt in order to evaluate their performance, while prompts involving occupations and animals provide more broad information such as facial engagement, gestures, and diversity of emotions, all of which can be measured to some degree using computer vision algorithms such as those described in Section 4.

Next, parents are directed to specify which child is about to play through the interface shown in Fig. 4. This is necessary to associate collected data from game sessions.
During a gameplay session, the caregiver flips the phone such that the screen is facing towards the child. The child will see a prompt and will act it out using gestures and expressions while being recorded with the phone’s front camera. When a child guesses correctly, the caregiver tilts the phone forward which awards them a point and changes the image prompt with the demographic information provided during the sign-up process. Moreover, identifying the current player is essential to support longitudinal progress tracking throughout multiple sessions. Parents can select from one of the children associated with their account, or select the Add a Player option at the bottom to add a new child. Future iterations of Guess What? will include automatic face-recognition features to detect the player without requiring manual selection.

After selecting a child, the orientation of the application automatically changes to landscape mode and the users are directed to an instructions page that describes the mechanics of the game. After they have reviewed the instructions, a 4-s countdown commences, and the users are redirected to the screen shown in Fig. 5. At this time, the phone screen will be directed outward towards the child, as they are recorded by the phone’s front camera. During the 90-s game session, the child will attempt to act out the prompt shown on the device while the parent guesses. After the parent guesses correctly, they will tilt the phone forward to indicate that the child has earned a point; the image will then change and the process will repeat until the timeframe has elapsed. The goal of the game is for the children to earn as many points as possible in the allotted time. If a child is struggling with a cue, the phone can be tilted backwards to skip the current prompt. Parents who prefer not to tilt the phone can also replace the current prompt by tapping on the image in the center of the game screen. The child is typically instructed not to speak or vocalize during the game, though this restriction may be lifted in some cases such as the animals deck.

After the game session is complete, users are presented with the screen shown in Fig. 6. Here, they can review the footage of their child and scroll through it to determine whether they wish to share this information with the research team or permanently delete it. Users who did not consent to share information about their child are not given this option, as the video is never recorded. If the user elects to share the footage, it is added to a local buffer and uploaded to a secure and encrypted IRB-approved Amazon Web Services S3 bucket in a background process when the user connects to WiFi.
To ensure user privacy, all captured videos during a game session are played back to the user immediately afterwards. At this time, the user can specify that they would like to delete the video. Alternatively, consenting participants can agree to share their video with us.

3.2 AWS

The entire backend infrastructure of Guess What? is based on Amazon Web Services technology. This includes Cognito for authentication and account creation, S3 Bucket for video storage, and DynamoDB for maintaining records of which children are associated with which parents, as well as their demographics. None of these databases are public, and they are all HIPAA compliant, IRB-approved, and fully compliant with the university’s High-Risk Application security standards.

The Amazon S3 bucket, used to store video data, maintains a separate directory for each user. Each directory consists of a series of sub-directories associated with a single game session. These folders consist of two files: the video file itself, as well as a log file that shows the timestamps at which the game began and ended, information about which prompts were shown and when, and which prompts were guessed correctly by the parent.

3.3 Gyroscope Signal Processing

During a game session, the user can proceed from one prompt to the next by tilting the phone fore and aft. If the phone is tilted in the forward direction, an annotation is
added to the log file to indicate that the prompt was guessed correctly by the parent. A backwards tilt simply skips the prompt, without awarding the child a point. The algorithm for detecting a phone tilt is based on the phone’s gyroscope sampled at 20 Hz. Equation 1 is the criteria used to detect a forward tilt. When the angular velocity in the x direction exceeds $70^\circ$ per second, with the constraint that movement

Fig. 7 This figure shows how a single video is processed. Each video frame is processed by a face tracker algorithm. Those frames in which a face is detected are processed using an emotion recognition algorithm. The distribution of emotions, as well as the percentage of face tracker failures, varies greatly between prompts, decks, and populations.
in the other dimensions is less than $40^\circ$ per second, a forward tilt is registered. This additional constraint is necessary to prevent random shaking to be interpreted as a tilt.

$$\omega_x < \frac{40^\circ}{\text{sec}}, \quad \omega_y > \frac{70^\circ}{\text{sec}}, \quad \omega_z < \frac{40^\circ}{\text{sec}}$$

(1)

The criteria to detect a backward tilt is very similar, with the exception that the angular velocity in the Y direction must be less than $-70^\circ$ per second. After each tilt, the tilt-detection algorithm is disabled for a period of 2 s to prevent multiple tilt events to be registered based on a single motion.

4 Algorithm Design

In this section, we describe the algorithms used to process videos acquired from Guess What? to better understand which prompts are most efficacious in provoking facial affect in participants with the more general purpose of revealing clear, quantifiable differences between children with ASD and those with no prior diagnosis from structured video.

Algorithm 1 Simplified data processing pipeline

```
1: procedure PROCESS_VIDEO(VideoData, GameData)
2:     while VideoFrame ≠ null do
3:         VideoFrame ← GetNextFrame()
4:         FaceFrame ← LocateFace(VideoFrame)
5:         if FaceFrame = null then
6:             TrackFailures ++
7:         else
8:             Label ← Classify(FaceFrame)
9:             if Label = Neutral() then
10:                NNeutral ++
11:             else
12:                NEmotion ++
13:         end if
14:     end while
15:     EmotePct = NEmotion / (NEmotion + NNeutral)
16: end procedure
```

4.1 General Architecture

Figure 7 shows the processing pipeline for a single video of a child playing one game session consisting of multiple prompts. As video is sampled at up to thirty frames per second, and facial expressions change at a far more gradual rate, the video can be
subsampled. These remaining frames, which are distributed evenly across the video, are processed using a face tracker algorithm which attempts to locate the face (if any) within the frame. When the child looks away from the camera during a game session, it is likely that the face tracker algorithm will fail. Therefore, the percentage of frames within a video in which the face tracker fails can be used as a general heuristic for a lack of facial engagement by the child during the game.

For those frames in which the face tracker is successful, a histogram of oriented gradients technique is used in conjunction with a linear SVM classifier to classify between eight classic emotions. These eight emotions include the seven universal emotions identified by Eckman et al. in [12], as well as a neutral category. For the purposes of our evaluation, we combine the non-neutral Ekman emotions into a single “emotive” class that is used in our subsequent analysis. The number of face tracker failures, percentage of neutral frames, and the percentage of emotive frames in which the face tracker was successful are all recorded and aggregated on a per-subject and per-deck basis. The primary metafeatures that we evaluate are summarized in Fig. 8. Note that in addition to the previous heuristics, we also evaluate the efficacy of using the child’s performance in the game as a feature—children with higher scores being presumably more efficient at conveying the prompts to their parents in a way that can be easily understood. A high-level summary of this algorithm for processing a video can be found in Algorithm 1. However, it should be noted that a meaningful analysis would be based on the aggregation of multiple videos rather than a single one.

![Figure 8](image)

Fig. 8 This figure shows the four primary metafeatures extracted from each video clip that are evaluated as a function of game category. We are interested in determining which categories produce the most clear differences between children with ASD and those with no prior diagnosis, as well as which categories are associated with high levels of emotive frames.
<table>
<thead>
<tr>
<th>Emotions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
</tr>
<tr>
<td>Sad</td>
</tr>
<tr>
<td>Surprised</td>
</tr>
<tr>
<td>Fear</td>
</tr>
<tr>
<td>Angry</td>
</tr>
<tr>
<td>Contempt</td>
</tr>
<tr>
<td>Disgust</td>
</tr>
<tr>
<td>Neutral</td>
</tr>
</tbody>
</table>

### 4.2 Emotion Classifier

Our implementation of the emotion classifier used in this work is based on prior work by Haber et al. in [19]. The emotions that can be automatically determined by our system are summarized in Table 1: happy, angry, neutral, surprised, contempt, fear, sad, and disgust. While we refer the readers to this previous work for a full description of the algorithm [19], we discuss it briefly in this section for completeness.

Our baseline system performs a classification in five steps: (1) face detection; (2) face tracking; (3) registration for rotation, translation, and scale; (4) lighting normalization; and (5) appearance-based feature extraction. Apple’s CIDetector is used for face detection and implements a face tracker which provides (up to) 23 landmark points around the face. An example of the output of the face tracker can be found in Fig. 9a as well as a failure of the face tracker in Fig. 9b. It should be noted that the face tracker implementation used in this system is suitable only for tracking direct faces, rather than peripheral or profile views.

In this work, we used the pre-trained classifier provided by Habe et al. in [19] without retraining. While we do not validate the accuracy of the emotion classifier on children in this paper, a performance analysis provided by Washington et al. in [36] indicated an accuracy of 60.4% for the base-model and 74.7% for the participant-specific model. Though not ideal, this performance is acceptable as the reported errors by Washington et al. are for all basic Ekman emotions while we focus on the simpler two-class problem of neutral and non-neutral classification.

### 5 Methods

A total of 13 children played played Guess What? on a Google Pixel. The following analysis is based on the 8 children within this cohort with a prior diagnosis for ASD (ages 8.5 years ± 1.85). To ensure consistency, the same individual, a member of our research staff, administered each game session. Participants played up to five games with the following decks: emoji, faces, animals, sports, and jobs. After these game sessions, WiFi was enabled and the data was synced to an Amazon S3 bucket. In one case, none of a subject’s videos were uploaded. In another case, only one video
This figure shows an example of a frame in which the face tracker has located a face and identified the associated landmark points shown in white. This frame will be processed to identify one of eight possible emotions. This frame is an example of a face tracker failure, in which the face tracking algorithm cannot locate a face within the frame because the subject is looking away from the camera. This frame is subsequently discarded and cannot be processed further in our existing implementation. It is through this heuristic that we estimate facial engagement during a game session from a subject was uploaded and the remainder failed to upload. Furthermore, there were several incidents in which the participants were not interested in continuing the game session beyond the first or second game. The eligibility criteria for this study required participants to be between the ages of 4 and 12 years old.

6 Results

In this section, we will describe the results of our analysis based on the data we have collected from eight subjects in our in-lab trial.

6.1 Results by Category

Table 2 shows the performance of participants in each of the categories, based on points-per-game and the amount of time spent per prompt. As expected, the jobs and sports categories are the most challenging as they require the most advanced figurative interpretation. By contrast, the easiest category was animals; children would often vocalize animal sounds rather than represent them using gestures and movements. The two remaining categories were emoji (showing emotion), and faces, which consisted of real photos of children showing emotion based on the standardized CAFE dataset of children expressing the Eckman emotions [12, 24]. Between these two, real photos of children appeared to be much easier for children to
Table 2  Average game statistics per category

<table>
<thead>
<tr>
<th>Emojis</th>
<th>Faces</th>
<th>Animals</th>
<th>Jobs</th>
<th>Sports</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Points per game</td>
<td>6</td>
<td>10</td>
<td>11</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Prompts shown</td>
<td>9</td>
<td>12</td>
<td>14</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Time per prompt (s)</td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>12</td>
<td>12</td>
</tr>
</tbody>
</table>

This table shows the average number of points earned by a child in each category—a heuristic for the difficulty of the category. Note that some of the younger participants did not play the more difficult categories, which inflated the average score of jobs and sports.

represent. This is perhaps because they were able to replicate the facial expression they saw in the photo: something that is more difficult with caricature-style emojis.

Developing an understanding of the difficulty of each prompt could be useful in the design of diagnostics and therapies that rely on progressive difficulty. Therefore, we show a list of the top fifteen most difficult and least difficult prompts in Table 3. Not surprisingly, the most and least difficult prompts were generally associated with the most and least difficult categories. However, there were several anomalies: the confusion and annoyed emoji was considered among the most difficult, while most other emoji were generally of average difficulty. A great amount of variation could be seen in sports, with archery, skateboarding, and biking being dramatically more difficult than basketball and baseball.

Figure 10 reveals which prompts differed most in time-to-guess from their category averages. The x-axis of this figure represents the number of seconds taken, on average, for participants to correctly guess the prompt. The y-axis indicates the difference between the time to guess the prompt and the class average, with a positive value indicating that the prompt was easier than others in its category. As shown, the dentist prompt was considerably more difficult than other occupations. Among sports, archery was more difficult than other categories in its class. By contrast, basketball and baseball were easier than other prompts in the sports category. Among animals, monkey and elephant were more challenging than animals with trivial vocalizations such as the pig, cow, and dog prompts. The monkey prompt in particular took 7.8 s longer than the class average. This was the largest overall observed discrepancy between a prompt and its class average.

There were few prompts within the faces category that varied in difficulty from the class average. By contrast, two prompts within emoji were anomalies: smile and confusion. Smile, being one of the most elementary emotions, was completed 4.8 s faster than the class average of 10 s. Confusion, by contrast, took an average of 14.2 s to express—4.2 s more time than the average emoji prompt.

Figure 11a shows the percentage of tracked frames in which a non-neutral emotion was recognized, across all game categories. The results are as expected; emotion-centric categories (faces and emoji) were associated with the highest level of non-neutral frames. By comparison, animals showed less emotion as children naturally preferred vocalizations. Sports, which emphasized gestures and movements,
Prompts that Differed in Difficulty From Category Averages

\[ \text{Diff. from Category Average (s)} \]

\[ \text{Average Time to Guess (s)} \]

**Fig. 10** The difficulties of prompts to others in the same category. The x-axis indicates the average number of seconds required to guess the prompt. The y-axis indicates the difference between this time and the class average. A positive value indicates that the prompt is more difficult than others within its category.

(a) Smile. (b) Baseball. (c) Basketball. (d) Elephant. (e) Monkey. (f) Confused (Emoji). (g) Archery. (h) Dentist

was associated with the lowest levels of emotive frames with 34% less frames than the highest category.

Figure 11b shows the percentage of face tracker failures associated with videos in each category. Recall that a face tracker failure always occurs when the subject looks away from the camera. Therefore, a higher percentage of failures in this category is associated with lower facial engagement. While we attribute the difference between emoji and faces (36% and 31% respectively) to low sample size, it is interesting that the vocalization-based animals category produced the highest amount of facial engagement among all categories; only 26% of frames from this category were associated with face tracker failures. While this category appeared to provide the most facial engagement, it also resulted in a low number of emotive frames.

Figure 12a shows the average percentage of frames within a video in which a non-neutral emotion was identified. Unlike Fig. 11, this graph illustrates the total percentage of emotive frames rather than the ratio of emotive frames to total tracked frames. Therefore, a lower percentage could be associated both with higher face tracker failures, a greater percentage of neutral frames, or some combination of these two factors. These results were generally in accordance with those from 11-A, with more emotive frames from emoji and faces compared to animals and sports.

Figure 12b shows the average percentage of frames within a video in which a neutral emotion was identified. Unlike Fig. 11, this graph indicates the total percentage of
Table 3  The fifteen most and least difficult prompts

<table>
<thead>
<tr>
<th>Time to guess (s)</th>
<th>Prompt</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.8</td>
<td>Dentist</td>
<td>Jobs</td>
</tr>
<tr>
<td>17.3</td>
<td>Archery</td>
<td>Sports</td>
</tr>
<tr>
<td>15.6</td>
<td>Skateboard</td>
<td>Sports</td>
</tr>
<tr>
<td>15.0</td>
<td>Biking</td>
<td>Sports</td>
</tr>
<tr>
<td>14.2</td>
<td>Confused</td>
<td>Emoji</td>
</tr>
<tr>
<td>14.0</td>
<td>Astronaut</td>
<td>Jobs</td>
</tr>
<tr>
<td>13.8</td>
<td>Monkey</td>
<td>Animals</td>
</tr>
<tr>
<td>13.5</td>
<td>Police</td>
<td>Jobs</td>
</tr>
<tr>
<td>13.1</td>
<td>Skiing</td>
<td>Sports</td>
</tr>
<tr>
<td>13.0</td>
<td>Elephant</td>
<td>Animals</td>
</tr>
<tr>
<td>12.6</td>
<td>Football</td>
<td>Sports</td>
</tr>
<tr>
<td>12.3</td>
<td>Pilot</td>
<td>Jobs</td>
</tr>
<tr>
<td>12.3</td>
<td>Doctor</td>
<td>Jobs</td>
</tr>
<tr>
<td>12.3</td>
<td>Love</td>
<td>Emoji</td>
</tr>
<tr>
<td>12.0</td>
<td>Ping-pong</td>
<td>Sports</td>
</tr>
<tr>
<td>6.1</td>
<td>Angry</td>
<td>Emoji</td>
</tr>
<tr>
<td>5.8</td>
<td>Happy</td>
<td>Faces</td>
</tr>
<tr>
<td>5.7</td>
<td>Baseball</td>
<td>Sports</td>
</tr>
<tr>
<td>5.4</td>
<td>Horse</td>
<td>Animals</td>
</tr>
<tr>
<td>5.4</td>
<td>Duck</td>
<td>Animals</td>
</tr>
<tr>
<td>5.2</td>
<td>Smile</td>
<td>Emoji</td>
</tr>
<tr>
<td>5.2</td>
<td>Cat</td>
<td>Animals</td>
</tr>
<tr>
<td>4.8</td>
<td>Cow</td>
<td>Animals</td>
</tr>
<tr>
<td>4.7</td>
<td>Surprise</td>
<td>Faces</td>
</tr>
<tr>
<td>4.7</td>
<td>Disgust</td>
<td>Faces</td>
</tr>
<tr>
<td>4.5</td>
<td>Neutral</td>
<td>Faces</td>
</tr>
<tr>
<td>4.1</td>
<td>Sheep</td>
<td>Animals</td>
</tr>
<tr>
<td>4.0</td>
<td>Lion</td>
<td>Animals</td>
</tr>
<tr>
<td>4.0</td>
<td>Frog</td>
<td>Animals</td>
</tr>
<tr>
<td>3.7</td>
<td>Pig</td>
<td>Animals</td>
</tr>
</tbody>
</table>

neutral frames rather than the ratio of emotive frames to total tracked frames. Here, the difference between categories was very low with the exception of the animals category, with a high number of neutral frames.

Figure 13 shows the number of points children earned, as a function of game category. The easiest category was the vocalization-centric animals, in which children earned 12 points throughout the 90-s game session. The sports category was the most difficult, with only six prompts earned. The emotion-centric categories, emoji and faces, were comparable, but children earned on average two more points for faces.
Fig. 11  Emotion classifier-based results for each category.  

<table>
<thead>
<tr>
<th>Category</th>
<th>Pct. of Emotive Frames Containing Face</th>
<th>Pct. of Face Tracker Failures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emoji</td>
<td>24%</td>
<td>36%</td>
</tr>
<tr>
<td>Faces</td>
<td>26%</td>
<td>31%</td>
</tr>
<tr>
<td>Animals</td>
<td>28%</td>
<td>24%</td>
</tr>
<tr>
<td>Sports</td>
<td>17%</td>
<td>27%</td>
</tr>
<tr>
<td>All</td>
<td>21%</td>
<td>35%</td>
</tr>
</tbody>
</table>

(a) The percentage of tracked frames with recognized non-neutral facial expressions shows that the *faces* category evokes the most emotion in participants.  

(b) *Animals*, a vocalization-centric category, was associated with the highest level of facial engagement despite having a low number of emotive frames.

This is in part because several of the *emoji* prompts, *confused* and *love*, had no analogue in the *faces* category and were likely more difficult to replicate exactly as these expressions were not displayed on human faces.

### 6.2 Summary of Findings

The primary contribution of our work is to present a system design and validation study demonstrating that a mobile charades-style game is a viable platform to obtain structured, emotive video from children with ASD. In this subsection, we summarize our findings and revisit the key questions posed in our introduction.

Fig. 12  Emotion classifier-based results for each category.  

<table>
<thead>
<tr>
<th>Category</th>
<th>Total Pct. of Emotive Frames</th>
<th>Total Pct. of Neutral Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emoji</td>
<td>15%</td>
<td>49%</td>
</tr>
<tr>
<td>Faces</td>
<td>17%</td>
<td>50%</td>
</tr>
<tr>
<td>Animals</td>
<td>13%</td>
<td>62%</td>
</tr>
<tr>
<td>Sports</td>
<td>13%</td>
<td>51%</td>
</tr>
<tr>
<td>All</td>
<td>17%</td>
<td>50%</td>
</tr>
</tbody>
</table>

(a) The percentage of frames containing non-neutral facial expressions, out of all possible frames in the video.  

(b) The percentage of frames containing neutral facial expressions, out of all possible frames in the video.
– Which categories elicit the most emotive and diverse facial expressions in children?
The category which produced the highest density of non-neutral frames in the associated gameplay video was *faces*, followed by the *emoji* category. Jobs and sports expectedly produce very little recognizable emotion. In terms of diversity within a video, the most efficacious category is *faces*: children were able to cover more prompts within a 90-s game session compared to *emoji*.

– Which categories and prompts are the most challenging, and which are the easiest?
The easiest category is animals, as children almost always expressed the prompt via vocalizations. By contrast, the jobs and sports categories are the most difficult. The easiest specific prompts are basic animals: pig, frog, lion, sheep, as well as several faces such as disgust and surprise. The most difficult prompts are sports outside of basketball, baseball, and football, as well as various occupations (dentist, astronaut).

– Are emoji representations of exaggerated emotions more suitable than photos of faces, for eliciting emotive facial expressions?
No: videos from the faces category produced a greater fraction of emotive frames in participants, with somewhat less emotion derived from emoji-based videos and fewer prompts covered within the game session. However, it should be noted that unlike the faces, the emojis were not from a standardized emotion-centric dataset and some prompts were therefore not associated with the seven Eckman universal emotions [12].

– Which category is associated with the highest level of facial engagement?
It is difficult to measure facial engagement directly with our existing system. However, despite producing relatively few non-neutral expressions, it was observed that the animals category was associated the lowest number of face tracker failures—a heuristic for facial engagement.
7 Limitations

There are several limitations associated with this study. First, the intention of the Guess What? platform was to crowdsource video capture for ASD research in at-home environments. However, the presented results were based on in-lab data collection administered by a single person. Data from at-home environments will be more varied in a number of ways including quality of footage, lighting conditions, willingness to aid the child, and demographics of participants. Thus, it will be necessary to attempt to replicate our findings from more varied home video.

Another limitation is the lack of a suitable control for this experiment. In the future, a much larger study will compare results from this game to distinguish between children with autism and a neurotypical control group. A third limitation is that we do not consider or measure audio-based feedback from children. This would be most advantageous in the animals category, as children would frequently make vocalizations associated with the animal shown in the prompt; the ability of a child to make the correct vocalization could be germane to the study of developmental delays in gamified systems. On that note, the fourth and final limitation of this study is the failure to automatically recognize gestures, which are more predominant in certain categories with lower facial engagement such as occupations and sports. While facial affect is the primary focus of this work, it is worth investigating which categories are most efficacious in producing gestures and if there are any implications on understanding autism.

8 Conclusion

In this paper, we present Guess What—a system for crowdsourcing video through a charades-style mobile game for applications in autism research. To validate the feasibility of this platform to acquire emotion-rich egocentric video, we conducted an in-lab study in which eight participants played the game. Generally, Guess What? was well-received as a gamified platform for data collection and research for ASD. Children were engaged and reacted positively to the different decks of prompts. The integration with Amazon Web Services provided reliable data capture, and videos were uploaded via background processes that required no direct interaction from the user.

Data obtained from vision-based emotion classifiers and aggregated statistics across game sessions provided insight about the efficacy of this platform in eliciting facial affect in participants that can be quantified and measured—important conclusions for the design of future therapies and home diagnostics for ASD. Specifically, we showed that game sessions in the faces category are associated with the most emotive facial expressions and the most prompts covered within a 90-s game session. It was also observed that the animals category produced the highest level of facial engagement by a considerable margin. Future work will incorporate audio and gesture recognition into the Guess What? platform, and these conclusions will be used in the design of future therapies and screening tools.
Funding Information  This study was supported by awards to D.P.W. by the National Institutes of Health (1R21HD091500-01 and 1R01EB025025-01). Additionally, we acknowledge support to D.P.W. from the Hartwell Foundation, the David and Lucile Packard Foundation Special Projects Grant, Beckman Center for Molecular and Genetic Medicine, Coulter Endowment Translational Research Grant, Berry Fellowship, Child Health Research Institute, Spectrum Pilot Program, and Thrasher Research Fund. The Dekeyser and Friends Foundation, the Mosbacher Family Fund for Autism Research, and Peter Sullivan provided additional funding.

Compliance with Ethical Standards

Conflict of interest  The authors declare that there is no conflict of interest.

Publisher’s Note  Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References