Short Papers

A Mobile Game for Automatic Emotion-Labeling of Images

Haik Kalantarian, Khaled Jedoui, Peter Washington, and Dennis P. Wall

***Abstract*—In this short paper, we describe challenges in the development of a mobile charades-style game for delivery of social training to children with autism spectrum disorder (ASD). Providing real-time feedback and adapting game difficulty in response to the child's performance necessitates the integration of emotion classifiers into the system. Due to the limited performance of existing emotion recognition platforms for children with ASD, we propose a novel technique to automatically extract emotion-labeled frames from video acquired from game sessions, which we hypothesize can be used to train new emotion classifiers to overcome these limitations. Our technique, which uses probability scores from three different classifiers and meta information from game sessions, correctly identified 83% of frames compared to a baseline of 51.6% from the best emotion classification API evaluated in this paper.**

***Index Terms*—Autism, crowdsourcing, emotion, mobile, domain adaptation, machine learning, deep learning.**

I. Introduction

Autism spectrum disorder (ASD) is a developmental disorder characterized by deficits in social communication and the presence of repetitive behaviors and interests [1]. The prevalence of this condition has increased in recent years, rising from an estimated 1-in-68 in 2010 to 1-in-59 in 2014 [1]. Although there is no cure for autism, multiple studies have demonstrated that applied behavioral analysis (ABA) therapy can improve developmental progress and social acuity if applied consistently from a young age [2]. The application of ABA therapy is customized to suit the child's deficits and needs, often including a method of teaching called discrete trial training (DTT) [2], [3].

Manuscript received May 29, 2018; revised August 1, 2018; accepted October 12, 2018. Date of publication October 22, 2018; date of current version June 16, 2020. This work was supported in part by awards to D. P. Wall by the National Institutes of Health (1R21HD091500-01 and 1R01EB025025-01) and in part by the Dekeyser and Friends Foundation, in part by the Mosbacher Family Fund for Autism Research, and in part by Peter Sullivan, in part by the Hartwell Foundation, in part by the David and Lucile Packard Foundation Special Projects Grant, in part by the Beckman Center for Molecular and Genetic Medicine, in part by the Coulter Endowment Translational Research Grant, in part by the Berry Fellowship, in part by the Child Health Research Institute, in part by the Spectrum Pilot Program, in part by the Stanford's Precision Health, in part by the Integrated Diagnostics Center (PHIND), and in part by the Stanford's Human Centered Artificial Intelligence Program. The work of H. Kalantarian was supported in part by the Thrasher Research Fund and in part by the Stanford NLM Clinical Data Science program (T-15LM007033-35) *(Haik Kalantarian and Khaled Jedoui contributed equally to this work.) (Corresponding author: Dennis P. Wall.)*

H. Kalantarian and D. P. Wall are with the School of Medicine and the Department of Pediatrics and Biomedical Data Science, Stanford University, Stanford, CA 94305 USA (e-mail: haik@stanford.edu; dpwall@stanford.edu).

K. Jedoui is with the Department of Mathematics and the School of Medicine, Stanford University, Stanford, CA 94305 USA (e-mail: thekej@stanford.edu).

P. Washington is with the Department of Bioengineering and the School of Medicine, Stanford University, Stanford, CA 94305 USA (e-mail: peter100@stanford.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TG.2018.2877325

A discrete trial is a unit of instruction delivered by the teacher to the child that lasts between 5 and 20 s, consisting of a prompt, response, reinforcement, and brief pause before the next trial [4]. Two areas in which DTT has been shown to be effective include 1) teaching new discriminations: the recognition of various cues often presented using flashcards; and 2) imitation: the ability to provide a response identical to the queue [4]. As difficulties in emotion recognition and expression are hallmarks of ASD [5], [6], DTT can be an integral component of a treatment program designed to address these deficits.

Caring for a child with autism can pose a significant financial burden on families [7]. This is in part due to interventions that require long hours of 1-on-1 therapy administered by trained specialists in increasingly short supply due to substantial growth in the incidence of this condition [1]. Alternatives that can ameliorate some of these challenges could be derived from digital and mobile tools. We are developing *Guess What?*: A mobile charades-style game that can potentially deliver a form of DTT to children at home through a social activity shared between the child, who must interpret and act out various emotive prompts shown on the screen, and the parent, who is tasked with guessing the emotion.

Presently, the caregiver is tasked with fulfilling the *reinforcement step* of DTT based on the fidelity of the child's imitation. Automatic image-based emotion recognition algorithms can supplement the reinforcement process by detecting if the child is emoting the correct prompt. This functionality can facilitate the development of additional game features that are integral aspects of ABA therapy: first, adapting the prompts to target the child's unique deficits [2], and, second, providing the appropriate visual feedback to guide the child toward the correct behavior without diminishing the challenge.

Most commercial emotion classifiers are trained on large databases of labeled images, such as the CIFAR-100, ImageNet [8], Cohn-Kanade Database [9], and Belfast-Induced Natural Emotion Databases [10]. While these datasets contain thousands of images, children are significantly underrepresented in these sources. Thus, classifiers trained on these databases are not optimized for vision-based autism research. This motivates the development of new approaches for scalable aggregation of emotive frames from children that can be used to design future classifiers and augment existing ones.

The primary contributions of this paper are as follows.

1) We present a preliminary version of our mobile charades-style game, *Guess What?, guesswhat*, as it is developed into a form of DTT for teaching emotion recognition and expression.
2) We describe the repurposing of the game for aggregation of emotive egocentric video for autism research.
3) We propose and evaluate two automatic labeling algorithms, which use probability scores from existing emotion classifiers and contextual meta-information to extract labeled frames from videos derived from these game sessions.

This paper is organized as follows. In Section II, we briefly cover related work in this area. In Section III, we describe the game design.

2475-1502 © 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

In Section IV, we present our algorithms. In Section V, we describe our experimental methods, followed by results in Section VI, and concluding remarks in Section VII.

II. Related Work

Meta classifiers are systems that consider the output of multiple classifiers as features, used to build another classifier for final class label assignment [12]. This technique has been applied to a variety of problem domains, such as grammar correction [13], news video classification, [12], and autism identification [14]. In [15], Chaibi proposed an ensemble-classification approach to detect emotions. Similarly, Perikos *et al.* proposed an ensemble-based method to detect emotion from textual data using bagging and boosting techniques [16]. Our approach leverages the success of these previous ensemble-based emotion recognition techniques by using classification scores from three classifiers combined with meta information from the game session to tag each frame with an emotion with a higher accuracy than any individual classifier could achieve.

In [17], Burmania *et al.* evaluated the ability of raters to correctly label data based on the inclusion of reference sets within the database with a predetermined ground truth. Sessions in which raters are performing poorly are therefore paused, which increases the accuracy of aggregated frames. This idea has also been explored in earlier works such as that of Le *et al.* [18]. These approaches are conceptually similar to our method, though we do not rely on human raters and instead characterize the systematic biases of individual classifiers in the determination of the final class label assignment.

In [19], Barsoum *et al.* used a deep-learning architecture to evaluate several manual labeling techniques to develop a framework in which scores from ten raters can be combined to generate a final label with highest accuracy. Similarly, Yu and Zhang [20] demonstrated that an ensemble of deep learning classifiers can significantly outperform a single classifier for facial emotion recognition. In contrast with these previous approaches, our method fuses classification confidence scores with game meta information rather than selecting the label with the maximum probability. By considering both per-class probabilities and *a priori* knowledge about the prompt shown at the time, labeling accuracy is significantly improved.

A novel approach for crowdsourcing labeled expressions is described in [21]. The authors propose an iPad puzzle game in which players are recorded through the device's front camera while periodically instructed to make various expressions. This paper bears similarity to our approach, however we target a younger audience and feature a social interplay between caregiver and child. Another crowdsourcing approach by Tuite and Kemelmacher [22] is the Meme Quiz: a game in which users are tasked with making an expression that the system attempts to recognize based on a continuously expanding training dataset. Results demonstrated statistically significant increases in classification accuracy over time despite an increasing numbers of classes. Unlike this platform, our system does not contain any online learning functionality in its current form. However, repeated game sessions will provide larger datasets that can be used to train a more robust emotion classifier offline.

Aside from crowdsourcing data, various educational and therapeutic games have been proposed in recent years. An example is Recovery Rapids: A gamified implementation of constraint-induced movement therapy for stroke rehabilitation in which the authors adapt therapy to the individual user in real time [23]. A similar work within the domain of autism research is described in [24]. Rather than acquiring labeled images, the authors propose an educational iPad game that targets deficits in emotion recognition and mimicry in children with developmental delay. During gameplay, the child is given a voice-instruction to imitate a face shown on the screen, which is presented beside the child's face acquired from the devices front camera. Our game, while similar, requires a caregiver to drive the experience due to the targeted age of participants.

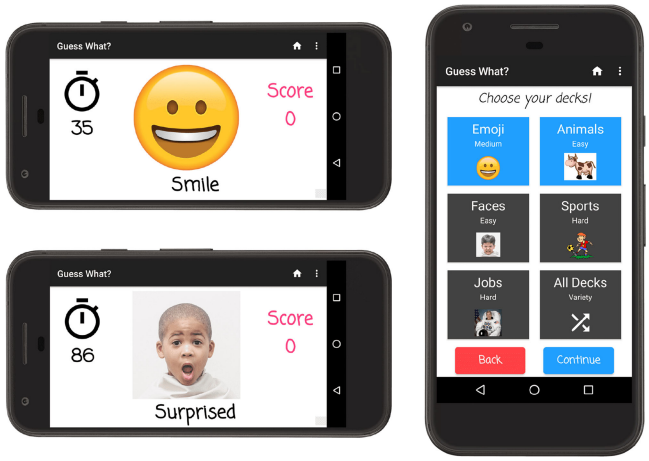

Fig. 1. In this mobile charades-style game, various prompts to the child during a 90-s game session. The parent attempts to guess the prompt as the child acts it out.

III. Game Design

Guess What? is a mobile game available for Android and iOS platforms [11], [35] designed to be a shared experience between the child, who attempts to enact the prompt shown on the screen through gestures and facial expressions, and the parent, who is tasked with guessing the word associated with the prompt during the 90 s game session. During each session, the parent holds the phone with the screen directed outward toward the child, who is recorded with the phone's front camera. This interplay, generally structured around the inversion-problem game described in [25], has the potential to provide a social, engaging, and educational experience for the child while providing structured video for researchers to develop a dataset of semilabeled emotion data.

While several categories of prompt are supported, the two most germane to emotion recognition and expression are *emoji*, showing exaggerated cartoon representations of emotive faces, and *faces*, which displays real photos of children. Examples of the main game screen when these two prompts are shown can be seen in Fig. 1, along with the main deck selection screen that allows users to select any combination of prompts to be shown during the 90-s game session.

When the child acknowledges the correct guess, or when the parent makes the determination that the prompt has been represented correctly based on *a priori* knowledge about the image shown, parents can change the prompt by tilting the phone forward to award a point. By tilting the phone backward, the prompt is skipped without awarding a point. Immediately thereafter, a new prompt is randomly selected until the 90 s have elapsed. The parent can determine the correct prompt by rotating the screen laterally to peek at the screen; the prompt only changes if a tilt in the longitudinal direction is detected. To reduce the likelihood of points being awarded accidentally, tilt detection is disabled for the first two seconds following the display of a new prompt.

After the game session, parents can review the footage and elect to share the data by uploading the video to an IRB-approved secure Amazon S3 bucket fully compliant with the Stanford University's High-Risk Application security standards. Meta information is included with the video, which describes the prompts shown, timing data, and the

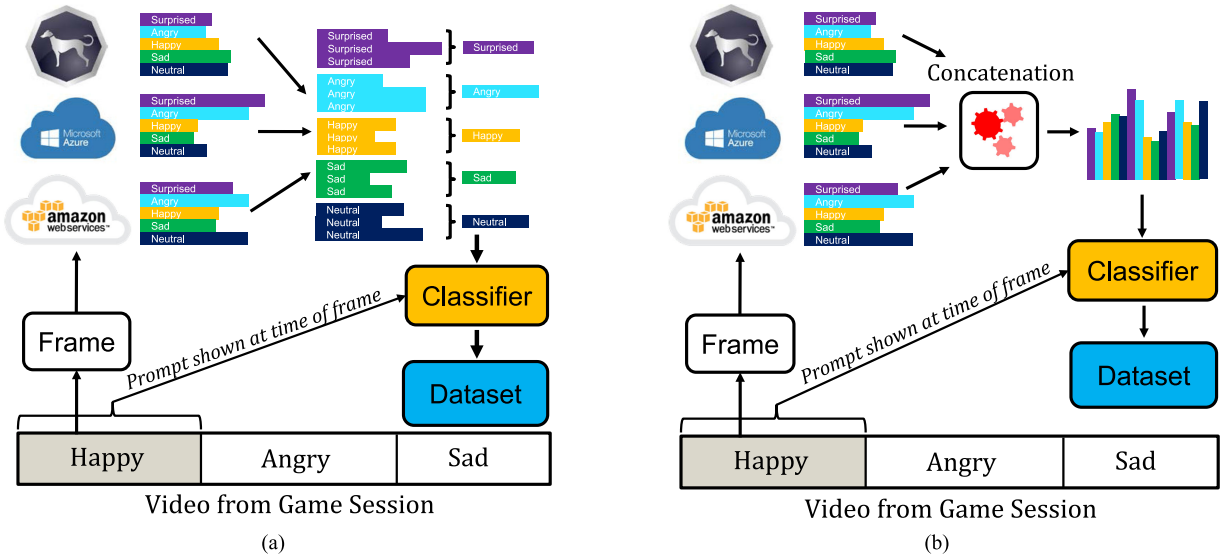

Fig. 2. (a) In the reduction-based feature processing algorithm, probability scores from each classifier are combined on a per-class basis before the final classification layer using *min*, *max*, or *average* functions. (b) In the aggregation-based feature processing algorithm, redundant probability scores from different classifiers are tagged as separate features.

number of points awarded. *Guess What?* is available for download online at: *guesswhat.stanford.edu*.

IV. Algorithms

The development of an emotion classifier that generalizes appropriately to children with ASD requires a substantial database of labeled images from this population. While videos derived from emotion-centric *Guess What?* game sessions are a natural choice to extract these frames, children are rarely able to consistently enact the correct prompts. Therefore, video segments in which images of a particular emotion are displayed cannot be used as a label for the frames contained therein. For example, preliminary results indicated that less than 25% of frames within regions of the video in which the *sad* emotion was shown matched the prompt. As manual frame-by-frame analysis is a tedious and error-prone process requiring trained raters, it is desirable to develop an automatic labeling approach that leverages meta information from the game, but that is not solely reliant on it. Therefore, an additional filtering step is necessary to remove invalid frames from video segments associated with a particular emotion.

A. Ensemble Learning for Emotion Classification

Commercial emotion recognition APIs, such as Azure Emotion API [26], Amazon Rekognition [27], and Sighthound [28], are convenient platforms for emotion recognition applications, and can be used to filter out frames within a region that are discordant with the prompt. However, these platforms showed remarkably poor performance on our datasets: no classifier identified over a third of frames in categories manually labeled as *sad*, *disgusted*, or *angry*. By combining the classification scores from multiple classifiers, we are able to effectively average out the nuances associated with each classifier, increasing the robustness of our classification. Two such approaches for combining classification confidence scores to label frames derived from game videos are shown in Fig. 2.

Fig. 2(a) shows how three emotion classifiers can be used to obtain three sets of classification confidence scores for each emotion. For example, Sighthound may report a 80% chance that the frame is *happy* with a 20% chance that it is *sad*, while AWS results may indicate 70%/30% probabilities. In the *reduction-based* architecture, classification confidence scores for each emotion are normalized and combined in three different ways: min, max, and average. A final classification layer then predicts the emotion associated with the image based on a reduced feature set consisting of a single probability score per emotion.

Fig. 2(b) shows the *aggregation-based* approach, in which probability scores from each emotion are not combined. Rather, probabilities are concatenated into a feature vector and all are used for prediction. In this approach, the final classification layer determines the best method to integrate duplicated probabilities from multiple classifiers into a final label assignment. This approach provided generally the higher performance compared to the *reduction-based method*; detailed results are presented in Section VI.

B. Last-Layer Classification for Label Assignment

In the final classification layer, the feature set generated from the outputs of three emotion classifiers is supplemented by the emotion of the prompt shown to the child at the time the frame was extracted to provide contextual meta information. This is based on the intuition that a frame is more likely to be *happy* if the prompt shown to the child at the time of the frame is related to a *happy* emotion. An additional classifier leverages this feature vector to assign a final emotion label to the frame. Random Forest [29] is a supervised ensemble learning method for classification that operates by training multiple decision tree classifiers on the dataset and predicting a class based on the number of votes by each decision tree. An advantage of Random Forest over other techniques is the use of bootstrap aggregation to improve predictive accuracy and control over-fitting.

Our classifier is trained for seven-class classification based on the maximum overlapping subset of a neutral class plus the six Ekman universal emotions [30] shared across the three classifiers: neutral, happy, sad, surprised, disgusted, scared, and angry. Two samples were designated as the minimum requirement to split an internal node, with at least one sample required for a leaf node. $\sqrt{N}$ features were considered when searching for the best split. For hyperparameters tuning, grid-search cross validation was used to determine the number of trees in each forest, as well as the maximum depth of each tree. The regression inputs were binned based on the techniques described in [31]. A total of

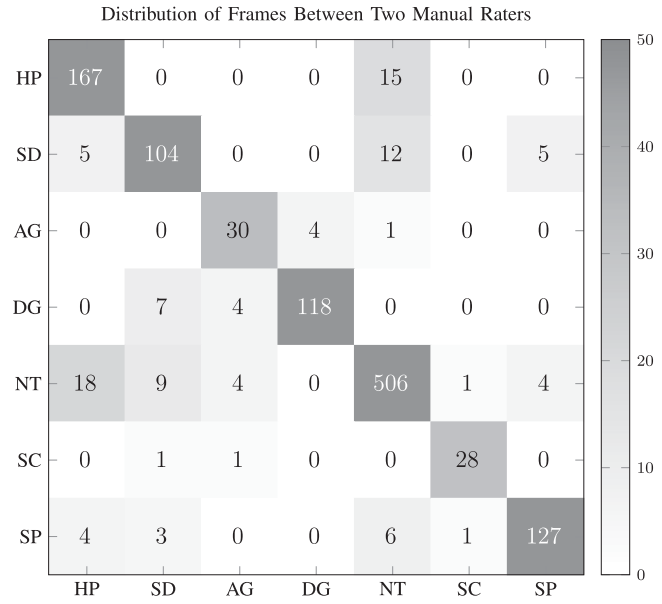

Fig. 3. Confusion matrix of the two raters assignments of frames into emotion categories. The abbreviations are: happy, sad, angry, disgust, neutral, scared, and surprised.

250 trees were used per forest, each with a maximum depth of 92. For maximum performance, final classification results are based on Leave One Out cross-validation (LOOCV): K-fold Cross Validation with K equal to the dataset size.

V. Experimental Methods

In this section, we describe the methods used to collect videos from *Guess What?*, manually label them to establish our ground truth, and leverage this dataset to validate our automatic labeling algorithms.

A. Data Collection

For algorithm evaluation, we drew from an internal dataset which included videos from eight children of ages 8.5 years $\pm$ 1.85 with a diagnosis of ASD. Participants played several *Guess What?* games in a single session administered by a member of our research staff on an Android phone. Due to the nonuniform incidence of autism between genders [32]–[34] and small sample size, all participants were boys. We analyzed video from games in one of the categories associated with affect, faces, producing a total of 8100 frames which were then subsampled to 5 frames per second.

B. Data Processing

To establish a ground truth, two raters manually assigned emotion labels to each frame in the selected videos based on the six Ekman universal emotions [30] with the addition of a *neutral* class. In cases, when no face could be located within the frame, or the frame was too blurry to discern, reviewers did not assign a label and the frame was excluded. To reduce the burden of manual annotation, the originally 30 FPS videos were subsampled to 5 FPS. From the selected videos within the *faces* category, a total of 1350 frames were manually labeled by the two raters. Frames were discarded in cases when the raters disagreed. This produced a total of 1080 frames from the original 1350. A confusion matrix showing the distribution of the rater's assignments can be seen in Fig. 3. The Cohen's Kappa statistic for interrater reliability, a metric that accounts for agreements due to chance, was 0.9. This indicates a high level of reliability between the two manual raters.

TABLE I
Baseline Accuracy With Existing Platforms

Technique	Accuracy	Weighted F1-Score
Azure Emotion	51.6%	0.62
AWS Rekognition	10.1%	0.12
Sighthound	9.2%	0.11

TABLE II
Ensemble Classification Without Meta Information

Technique	Features	Accuracy	Weighted F1-Score
Minimum probability	7	47.5%	0.36
Maximum probability	7	66.8%	0.64
Average probability	7	66.4%	0.64
All probabilities	21	76.6%	0.75

VI. Results

In this section, we demonstrate our algorithm's performance in automatic labeled frame extraction based on the ground truth established by two manual raters.

A. Baseline Labeling Accuracy

To establish a baseline to compare performance, we ran the 1080 derived frames through three Emotion Classification APIs using no meta information and selected the class with the highest probability as the final label assignment. As shown in Table I, results were poor; the highest accuracy was achieved by Azure Emotion [27] at 51.6%. The Sighthound API [27] produced the lowest accuracy at 9.2%, followed by AWS Rekognition [26] at 10.1%. Note that the weighted F1-scores are significantly higher than accuracy based on total percentage of correctly classified instances. This indicates that these platforms are tuned to correctly recognize the most common emotions such as *happy* and *neutral* to the detriment of other less common emotions. These results preclude the integration of these emotion classifiers into *Guess What?* and necessitate novel methods to automatically label frames using ensemble-techniques and game context.

B. Ensemble-Classifier Results: No Meta Information

The performance of the ensemble-labeling technique with no meta information is shown in Table II. Specifically, this approach is based on Random-Forest classification of a feature set of confidence scores derived from three emotion classifiers, but does not include the prompt shown at the time the frame was extracted as a feature. These results indicate that the *aggregation*-based technique shown in Fig. 2(b) is a better labeling approach than the reduction-based techniques shown in Fig. 2(a) for overall classification accuracy. Furthermore, we can conclude that both *aggregation* and *reduction*-based techniques outperform all three commercial emotion recognition APIs even without the inclusion of game context.

C. Ensemble-Classifier Results: With Meta Information

Results for ensemble-based approaches that use probability scores from all three classifiers in addition to game meta information are shown in Table III. The *aggregation* approach, which treats scores from each classifier as separate features, was associated with the highest ac-

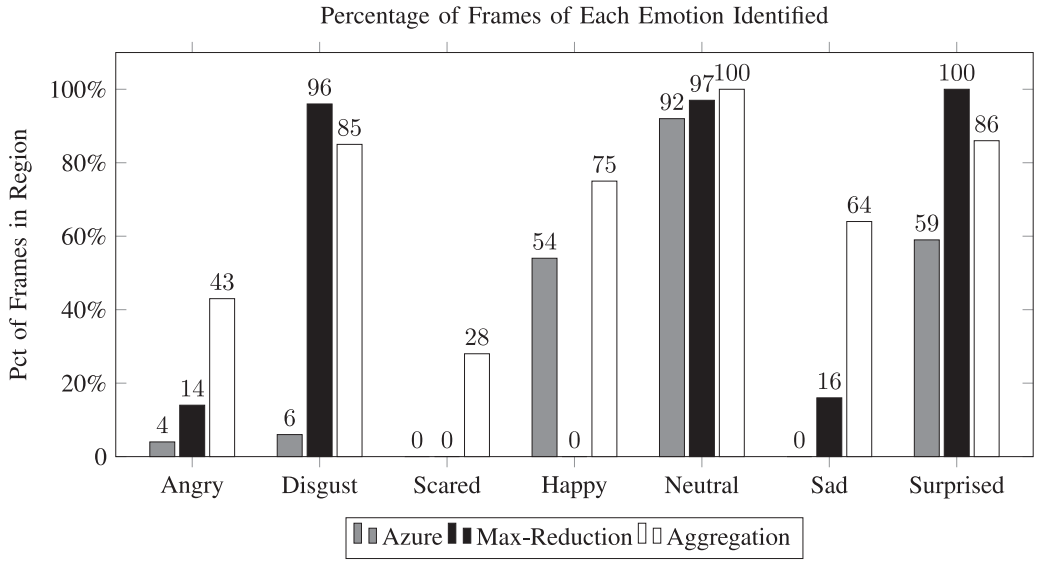

Fig. 4. Comparison of Azure Emotion API's performance with our max-reduction and aggregation techniques on a per-class basis shows that our algorithm can correctly label the majority of frames in every category except *scared* and *angry*, and our aggregation-based method outperforms the best of the three commercial emotion recognition APIs in every category.

TABLE III
ENSEMBLE CLASSIFICATION WITH META INFORMATION

Technique	Features	Accuracy	Weighted F1-Score
Minimum probability	8	61.0%	0.54
Maximum probability	8	78.3%	0.77
Average probability	8	77.5%	0.77
All probabilities	22	83.4%	0.84

TABLE IV
OVERALL ACCURACY COMPARISON

Technique	Accuracy	Weighted F1-Score
Baseline (best)	51.6%	0.62
Ensemble without meta (best)	76.6%	0.75
Ensemble with meta (best)	83.4%	0.84

curacy, at 83.4%: a significant improvement over the best emotion classification API's 62.6% accuracy. While all *reduction*-based approaches had lower accuracy, the performance of the max-probability and average-probability approaches were similar at 78.3% and 77.6%, respectively. As before, the minimum-probability approach had the lowest accuracy at 61.0%.

D. Comparison of Methods

Table IV shows the best possible accuracy achieved by the baseline classifier with no meta information (Azure), the ensemble technique with no meta information, and the technique with the highest accuracy: the ensemble classifier that includes game-meta information, which correctly identified 83.4% of frames.

E. Emotion-Specific Results

Fig. 4 shows the percentage of frames within each region that were correctly identified by the best commercial API evaluated (Azure), the best reduction technique (max), and the aggregation-based technique. Specifically, the reported accuracy for each emotion represents the percentage of identified frames of that class within game periods in which the image associated with that emotion were shown. It should be noted that the Azure classifier operates solely on the data frame and does not take into consideration any meta information from the game session.

Several conclusions can be drawn from this data. First, per-class results indicate that the Azure API performance was highly inconsistent between different emotions: the system performed well on *neutral* and *happy* frames but performed poorly on others. Second, results show that the aggregation-based technique does not outperform the max-reduction algorithm for every emotion: max-reduction identified a greater percentage of frames in *disgust* and *surprise* categories. Finally, even the best of the two labeling techniques still could not correctly identify a majority of frames associated with *angry* and *scared* classes; this may be a consequence of a limited training set and will be addressed in future work.

VII. CONCLUSION

In this paper, we have presented a mobile charades-style game in active development, designed to deliver emotion-recognition training to children with ASD. We describe how this platform can be used to derive emotion-rich egocentric video for processing with automatic labeling algorithms that fuse game-meta information with probability scores to predict emotion with a higher accuracy than commercial emotion recognition APIs. In future work, the labeled frames extracted from these videos will be used to train an emotion classifier that generalizes to children with ASD, which will be integrated into the game to provide reinforcement as social deficits are addressed through gameplay, for example, via at-home DTT.

REFERENCES

[1] J. Baio *et al.*, "Prevalence of autism spectrum disorder among children aged 8 years," *MMWR Surveillance Summaries*, vol. 67, no. 6, pp. 1–23, 2018.

[2] R. M. Foxx, "Applied behavior analysis treatment of autism: The state of the art," *Child Adolescent Psychiatric Clin.*, vol. 17, no. 4, pp. 821–834, 2008.

[3] N. R. Council *et al.*, *Educating Children With Autism*. Washington, DC, USA: National Academies Press, 2001.
[4] T. Smith, "Discrete trial training in the treatment of autism," *Focus Autism Other Developmental Disabilities*, vol. 16, no. 2, pp. 86–92, 2001.
[5] M. B. Harms, A. Martin, and G. L. Wallace, "Facial emotion recognition in autism spectrum disorders: A review of behavioral and neuroimaging studies," *Neuropsychol. Rev.*, vol. 20, no. 3, pp. 290–322, 2010.
[6] H. Macdonald *et al.*, "Recognition and expression of emotional cues by autistic and normal adults," *J. Child Psychol. Psychiatry*, vol. 30, no. 6, pp. 865–877, 1989.
[7] C. Horlin, M. Falkmer, R. Parsons, M. A. Albrecht, and T. Falkmer, "The cost of autism spectrum disorders," *PloS One*, vol. 9, no. 9, 2014, Art. no. e106552.
[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[9] J. Cohn *et al.*, "Cohn-Kanade AU-coded facial expression database," Pittsburgh University, Pittsburgh, PA, USA, 1999.
[10] E. Douglas-Cowie , R. Cowie, and M. Schröder, "A new emotion database: Considerations, sources and scope," in *Proc. ISCA Tut. Res. Workshop Speech Emotion*, 2000, pp. 39–44.
[11] 2018. [Online]. Available: guesswhat.stanford.edu
[12] W.-H. Lin and A. Hauptmann, "News video classification using SVM-based multimodal classifiers and combination strategies," in *Proc. 10th ACM Int. Conf. Multimedia*, 2002, pp. 323–326.
[13] M. Gamon, "Using mostly native data to correct errors in learners' writing: A meta-classifier approach," in *Proc. Human Lang. Technol., Annu. Conf. North Amer. Chapter Assoc. Comput. Linguist.*, 2010, pp. 163–171.
[14] E. I. Papageorgiou and A. Kannappan, "Fuzzy cognitive map ensemble learning paradigm to solve classification problems: Application to autism identification," *Appl. Soft Comput.*, vol. 12, no. 12, pp. 3798–3809, 2012.
[15] M. W. Chaibi, "An ensemble classifiers approach for emotion classification," in *Proc. Int. Conf. Intell. Interact. Multimedia Syst. Serv.*, 2017, pp. 99–108.
[16] I. Perikos and I. Hatzilygeroudis, "A classifier ensemble approach to detect emotions polarity in social media," in *Proc. Int. Conf. Web Inf. Syst. Technol.*, 2016, pp. 363–370.
[17] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Trans. Affective Comput.*, vol. 7, no. 4, pp. 374–388, 2016.
[18] J. Le, A. Edmonds, V. Hester, and L. Biewald, "Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution," in *Proc. SIGIR Workshop Crowdsourcing Search Eval.*, 2010, vol. 2126.
[19] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, 2016, pp. 279–283.
[20] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2015, pp. 435–442.
[21] C. T. Tan, H. Sapkota, and D. Rosser, "Befaced: A casual game to crowdsource facial expressions in the wild," in *Proc. CHI'14 Extended Abstract Human Factors Comput. Syst.*, 2014, pp. 491–494.
[22] K. Tuite and I. Kemelmacher, "The meme quiz: A facial expression game combining human agency and machine involvement," in *Proc. Int. Conf. Found. Digit. Games*, 2015.
[23] D. Maung *et al.*, "Development of recovery rapids-a game for cost effective stroke therapy," in *Proc. Int. Conf. Found. Digit. Games*, 2014.
[24] N. Harrold, C. T. Tan, and D. Rosser, "Towards an expression recognition game to assist the emotional development of children with autism spectrum disorders," in *Proc. Workshop SIGGRAPH Asia*, 2012, pp. 33–37.
[25] L. Von Ahn and L. Dabbish, "Designing games with a purpose," *Commun. ACM*, vol. 51, no. 8, pp. 58–67, 2008.
[26] 2018. [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/emotion/
[27] 2018. [Online]. Available: https://aws.amazon.com/rekognition/
[28] 2018. [Online]. Available: https://www.sighthound.com/products/cloud
[29] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324
[30] P. Ekman *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion," *J. Personality Soc. Psychol.*, vol. 53, no. 4, pp. 712–717, 1987.
[31] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, vol. 1. New York, NY, USA: Springer, 2001.
[32] "Autism society: What is autism?," 2016. http://www.autism-society.org/what-is/. Accessed on: Oct. 30, 2017.
[33] G. Dawson, "Early behavioral intervention, brain plasticity, and the prevention of autism spectrum disorder," *Development Psychopathol.*, vol. 20, no. 3, pp. 775–803, 2008.
[34] G. Dawson and R. Bernier, "A quarter century of progress on the early detection and treatment of autism spectrum disorder," *Development Psychopathol.*, vol. 25, pt. 2, no. 4, pp. 1455–1472, 2013.
[35] H. Kalantarian *et al.*, "Guess what?," *J. Healthcare Inf. Res.*, Springer International Publishing, 2018, pp. 1–24.