Infant Behavior and Development 54 (2019) 1–12

Contents lists available at ScienceDirect

Infant Behavior and Development

journal homepage: www.elsevier.com/locate/inbede

Full length article

Using automatic face analysis to score infant behaviour from video collected online

Brea Chouinard[a], Kimberly Scott[b], Rhodri Cusack[a,*]

[a] Trinity College Dublin, Dublin, Ireland
[b] Massachusetts Institute of Technology, Cambridge, MA, USA

ARTICLE INFO

Keywords:
Face detection
Machine vision
Looking time
Preferential looking
Webcam

ABSTRACT

Online testing of infants by recording video with a webcam has the potential to improve the replicability of developmental studies by facilitating larger sample sizes and by allowing methods (including recruitment) to be specified in code. However, the recorded video still needs to be manually scored. This labour-intensive process puts downward pressure on sample sizes and requires subjective judgements that may not be reproducible in a different laboratory. Here we present the first fully automatic pipeline, using a face analysis software-as-a-service and a discriminant-analysis classifier to score infant videos acquired online. We compare human and machine performance for looking time and preferential looking paradigms; machine performance demonstrates a promising proof of principle for looking time and is above chance in classifying preferential looking. Additionally, we studied the characteristics of the video and the child that influenced automated scoring, so that future studies can acquire data that maximises the performance of automatic gaze coding and/or focus on improving automatic coding for particularly challenging data. We believe this technology has great promise for developmental science.

1. Introduction

The value of open science has been recognised since the time of the Ancient Greeks (Resnik, 2006). Central to the method is that studies are transparently described so that they can be repeated by other experimenters, and that they have sufficient power and generalisability that their findings will typically replicate. Open science has proven enormously successful, but countervailing pressures have encouraged scientists to be lax in their description of methodology and to publish work that is insufficiently powered to replicate. As a result, many fields have been weakened (Button et al., 2013; Munafò et al., 2017; Pashler & Harris, 2012), particularly in the challenging areas for which theory remains underdetermined and data are time-consuming or expensive to acquire, including psychology (Aarts et al., 2015) and developmental science (Frank et al., 2017). We propose that internet-based testing can facilitate openness in an important area of developmental science, the study of the behaviour of infants, by providing transparent methods that are easily shared; allowing different laboratories to recruit from the same diverse population of participants; and increasing statistical power through larger sample sizes. In this paper, we evaluate for the first time a fully automated, transparent and replicable pipeline for acquisition and analysis of infant behavioural experiments.

Online services such as Amazon's Mechanical Turk (MTurk; mturk.com) or Prolific (prolific.ac) have been shown to allow rapid and inexpensive recruitment of participants for research (Buhrmester, Kwang, & Gosling, 2011). Early online studies targeted adults and used (Lefever, Dal, & Matthíasdóttir, 2007; Lonsdale, Hodge, & Rose, 2006) or evaluated (Granello & Wheaton, 2004; Buhrmester

* Corresponding author at: Trinity College Institute of Neuroscience, Trinity College Dublin, 2 College Green, Dublin 2, Ireland.
E-mail address: rhodri@cusacklab.org (R. Cusack).

https://doi.org/10.1016/j.infbeh.2018.11.004
Received 2 May 2018; Received in revised form 30 September 2018; Accepted 19 November 2018
Available online 30 November 2018
0163-6383/

et al., 2011; Paolacci, Chandler, & Ipeirotis, 2010; Rand, 2012; Riva, Teruzzi, & Anolli, 2003) web-based surveys or questionnaires. Online data collection was found to be cheaper than paper-based methods, facilitating larger sample sizes (Ebert, Huibers, Christensen, & Christensen, 2018) and had high test-retest reliability (Buhrmester et al., 2011). The prevalence of webcams has made it possible to move beyond surveys and to acquire video data (McDuff, El Kaliouby, & Picard, 2015). This has recently been used for the first online studies of infant behaviour (Scott & Schulz, 2017; Scott, Chu, & Schulz, 2017; Semmelmann, Hönekopp, & Weigelt, 2017; Tran, Cabral, Patel, & Cusack, 2017).

Online infant testing was less effortful and less time consuming for both scientists and the parents of the participants. However, to analyse the video data, manual coding was still necessary (Tran et al., 2017). This is extremely time consuming, which places a downwards pressure on sample sizes, and is subjective, which may introduce systematic differences between laboratories and reduce replicability. In the current study, we aimed to establish for the first time a fully-automated pipeline for infant behavioural experiments, with scoring of video samples using automatic face recognition with the cloud-based face recognition service provided by Amazon Web Services, Rekognition Video (Amazon, 2017). It uses a deep learning algorithm to locate faces, estimate their age and head orientation in three dimensions, and to identify landmarks including the pupils.

To evaluate this tool, we used pre-existing manually-coded data and compared the performance of Amazon Rekognition to the human ratings. We evaluated performance in two common developmental paradigms. In the looking time paradigm, the amount of the time that an infant looks at a stimulus is measured. In the preferential looking paradigm, two stimuli are presented at different positions and the amount of the time looking to each is measured. To code these tasks, the system must locate the infant face in the image and determine where its gaze is directed. In addition to measuring performance, we quantified characteristics of the video and participant that might influence the success of machine scoring, to provide a guide on what aspects of data acquisition and processing could be improved in future studies. Any of the benefits found in this automated pipeline would also be applicable to video data that are gathered through in-person visits.

2. Experiment 1: detecting looks to and away from the screen

Looking time paradigms are common in developmental research, and can be used to measure many things, including intentionality (Hamlin, Wynn, & Bloom, 2007; Woodward, 1998), emotion (Montague & Walker-Andrews, 2001), and speech preference (Cooper, 1990; Maye, Werker, & Gerken, 2002). The dependent variable is typically either the total amount of time the infant spends looking at the stimulus within a fixed interval, or how long the infant initially looks at the stimulus before the first look-away of some predefined minimum duration. In our studies, a stimulus of interest was presented on the computer monitor, and we evaluated how well a machine algorithm classified the infants as looking vs. not looking at the screen.

2.1. Methods

2.1.1. Video recordings

Data for Experiment 1 were obtained from two pre-existing studies in which videos had been laboriously tagged by human raters, the Oneshot study from Lookit (Scott & Schulz, 2017) and the study by Tran et al. (2017). The numbers and durations of the video samples for each sample are shown in Table 1.

The Oneshot study was part of a set of initial studies conducted to evaluate the viability of the Lookit online developmental lab. Lookit is run by scientists at Massachusetts Institute of Technology and allows parents and children to participate in developmental research via webcam (Scott & Schulz, 2017). In the Oneshot study, infants aged 11 to 17 months sat on a parent's lap, and videos were played on the screen, with the child's looking behavior captured via webcam. The data were made available through the Open Science Framework (https://osf.io/mbcu2/), in collections based on the video privacy level selected by the parent. For the current study, a total of 46 Oneshot videos were processed, 19 from the publicity data set ("OSP," where consent is available for publicity and educational purposes) and 27 from the scientific data set ("OSS," videos to be used for scientific purposes only). We processed these collections separately to facilitate work building on the results presented here; a reader using only the OSP videos can still replicate the relevant results. Each Oneshot video represented eight 20-s trials (four warm-up and four test) from a single session. The videos had previously been tagged independently by two human raters for looking time using Vcode (Hagedorn, Hailpern, & Karahalios, 2008), resulting in a record of the start and end of each look to the screen.

We also used 14 videos from Tran et al. (2017), which were acquired through Amazon MTurk. Infants ranged in age from 5 to 8

Table 1
Demographic information for study samples in Experiment 1 (Looking Time; left panel) and Experiment 2 (Preferential Looking; right panel) reported by site: Oneshot Publicity Level (OSP), Oneshot Scientific Level (OSS), Tran et al., 2017 (Tran); NovelVerbs Publicity (NVP); NovelVerbs Scientific (NVS).

	Looking Time			Preferential Looking	
	OSP	OSS	Tran	NVP	NVS
Number of videos	19	27	14	42	58
Range of seconds of video tagged by human raters (M, SD)	156-171 (164, 4)	161-494 (177, 63)	588-768 (655, 59)	141-170 (157, 6)	127-463 (163, 41)

months, and were seated on a parent's lap. They viewed 9–13 minute clips of children's television programs, and their responses were recorded by webcam. One human rater had previously annotated the video samples using Anvil (Kipp, 2014).

2.1.2. Face detection and pre-processing

We used an online face recognition service, Amazon Rekognition (Amazon, 2017). We chose this platform because it provides rich information on the location, orientation, and configuration of the face, and because it is cloud-based, readily available, and highly scalable. Its cost at the time of writing is $0.10 per minute of video analysed (https://aws.amazon.com/rekognition/pricing).

The 60 videos were uploaded to Amazon S3 and processed with Rekognition using code written in Python 3.63, with the Boto 3 module to interface with Amazon Web Services, and scikit-learn for discriminant analysis (www.github.com/rhodricusack/aws_video). Rekognition executes asynchronously, allowing batch processing of multiple videos simultaneously. On completion, it was configured to broadcast a notification to Amazon SNS, containing a JSON file describing the results, which was then passed to an Amazon SQS queue. Rekognition processes video in time segments, detecting faces in each segment. For each face it yielded: the detection time; a bounding box; an estimate of its age; a "Confidence" variable indicating the percentage of confidence in the presence of a face; and "Brightness" and "Sharpness" variables with higher percentages representing brighter or sharper face images. Rekognition's time segments were found to vary from movie to movie in the range 83–266 ms. The results were downloaded from Amazon SQS to our lab server for processing. To allow manual inspection, the results were superimposed onto the original videos.

To distinguish the infants from their parents, who were typically also in the video frame, we filtered for faces that were estimated as younger than 10 years of age, chosen as it was approximately half way between the age of the infants and young parents. The first major challenge for the face recognition algorithm, is to locate the faces, and correctly identify the infant faces. As all frames/almost all frames contained one (and only one) infant face, and the adult faces were never classified as infant faces, to assess the performance of the classifier to detect the infant face, we calculated for each video the proportion of frames in which exactly one infant face was found ("PropOneFace"). PropOneFace serves as a summary metric of success of video preprocessing, indicating the ability of the automated system to find the relevant region of the video for coding. Although we expect almost all videos to in fact contain exactly one infant face almost all of the time, the "correct" value of PropOneFace may be less than one when a child's face is out of frame for some portions of the video or when a sibling is present in the background.

2.1.3. Classification

To provide a basis from which to estimate the infant's gaze orientation, the following features were extracted: head orientation in three dimensions (pitch, yaw and roll); the position of the left eye, right eye, left pupil, and right pupil (as coordinates in the image); whether the eyes were open or closed; and the confidence of the eye openness. While Rekognition can detect the presence of a face and the location of these various "landmarks," it does not estimate the direction or target of gaze. In principle, this is a difficult problem that requires rough reconstruction of a 3-dimensional scene – including the unseen screen that the subject is viewing – from a webcam video frame. However, for the purposes of looking behaviour coding we often require only very coarse information – in this case, whether the child is looking or not. We sought to evaluate here whether this information could typically be recovered simply from the landmark position estimates.

One major challenge for generalization is that children's position varies relative to the video monitor and webcam, with systematic differences across subjects (due to varying setups) and movement within subjects. Users' webcams are not always horizontally centered relative to the screen, so even if we could reliably detect looking "head on" versus to the side, this would not directly yield the gaze target on the screen. For this initial approach, we simply zero-centered the orientation (pitch, yaw, and roll of the head) and position of the eyes and pupils for each participant across time, making the crude assumption that the infant's gaze is centered on average in the center of the screen, and that this is reflected in the mean positions of the features. This necessarily reduces the information available to the classifier, as we expect that typical conditions do in fact violate this assumption – e.g., if a child tends to look away to one direction rather than the other (for instance, towards the rest of the room instead of towards a wall) or displays a genuine preference for one side of the screen rather than the other in a preferential looking paradigm. Therefore, we emphasize that the expected performance will not be human-level coding, but *any* success following this transformation – that is, an ability to tell which looks are *closer* to center or which are *more* to the left or right for this child - represents substantial promise for future approaches that incorporate either per-child manually coded training data or a calibration phase.

Within each experimental dataset, a discriminant analysis classifier was trained to associate the pre-processed features with the manually coded state (looking vs. not looking). When more than one human coder was available, the median value at each time point was used. To train and test the classifier, we used leave-one-subject-out cross validation - for each set of N videos we trained the classifier N times, leaving out one subject each time and testing performance on that subject. Success in classifying data therefore represents passing a high bar: generalization to a new subject and video, with *no* use of training data from that video, despite known variation in webcam and screen positioning across subjects. We used a quadratic discriminant classifier because we expected input features to be non-linearly associated with whether the child was looking – for example, a pupil positioned far to the left (large negative x coordinate) or far to the right (large positive x coordinate) would indicate the infant was not looking, but a pupil close to center (small x coordinate) would indicate looking. Classifier performance was tested by comparing the classifier's predictions with the manually coded values in the left-out subject, using a receiver operating characteristic (ROC) curve and d-prime from signal detection theory.

2.1.4. Manual rating of quality of images

On a large subset of videos, we also acquired a set of human ratings of the video characteristics, by rating two frames that were

Table 2
Information for study samples in Experiment 1 (Looking Time; left panel) and Experiment 2 (Preferential Looking; right panel) reported by site; Oneshot Publicity Level (OSP), Oneshot Scientific Level (OSS), Tran et al., 2017 (Tran); NovelVerbs Publicity (NVP); NovelVerbs Scientific (NVS).

	Looking Time			Preferential Looking	
	OSP	OSS	Tran	NVP	NVS
Range of PropOneFace (M, SD)	.01-.98 (.60, .26)	.02-.98 (.74, .22)	.17-.97 (.60, .26)	.07-.96 (.70, .24)	.03-.93 (.69, .22)
Average d-prime (SD)	1.09 (.66)	1.47 (.60)	1.70 (.44)	0.71 (.70)	0.79 (.33)

randomly selected from each video. These ratings, performed by the first author, aimed to quantify characteristics of the video (lighting, resolution, horizontal and vertical position of light source) and child (iris size and color, eye shape, and whether the face was fully in the frame) that might influence the ability of the automatic algorithm to detect the faces and to judge gaze direction (see supplementary Table S1).

2.2. Results

2.2.1. Were infant faces automatically detected?

Rekognition succeeded in detecting a single infant face in every video at least some the time. PropOneFace, the proportion of frames coded by human raters in which a single infant face was detected by the classifier, is summarized in Table 2 and illustrated in Fig. 1a. A single infant face was detected for approximately 70% of frames on average, but there was substantial variation across videos, with PropOneFace ranging from 0.01 (single infant face almost never detected) to 0.98 (almost always detected).

2.2.2. Did the classifier correctly detect looking towards the screen?

We calculated the sensitivity index, d-prime, for each video to quantify the sensitivity of the automated tagging. The classification task can be formulated as detecting looks towards the screen: for frames where the human coder says the child is looking, if the classifier agrees that is a "hit" and if it says the child is not looking that is a "miss." For frames where the human coder says the child is not looking, if the automatic system agrees that is a "correct rejection" and if the automatic system says the child *is* looking that is a "false alarm." d-prime is calculated z(H) – z(FA), and a higher d-prime indicates that the hit rate (hits divided by frames where the child was looking) is higher than the false alarm rate (false alarms divided by frames where the child wasn't looking), thus better sensitivity or classifier accuracy. Sensitivity was far above chance (d-prime = 0) across experiments, 59 positive and 1 negative (Sign test $p < 0.001$). The ROC curves in Fig. 2a and our d-prime values (Table 1) indicate that the classifier performed best when detecting looking vs. no looking in the Tran data (green dots closest to the most bowed line) with acceptable, above chance performance on the OSP and OSS data (orange and blue dots, respectively, spread between two bowed curves and far from the diagonal line, which represents chance). A one-way ANOVA indicated a significant difference between the subsamples, $F_{2,59} = 4.667$, $p = .013$, and follow-up two-tailed *t*-tests indicated that Tran sensitivity was better than OSP, $t(31) = 2.947$, $p = .006$, but not OSS $t(39) = 1.469$, $p = .15$, and with no significant difference between OSS and OSP, $t(44) = 1.919$, $p = .062$. Pooling the three datasets, there was a significant correlation between PropOneFace and d-prime ($r = .37$, $p = .004$, two-tailed) indicating that when the proportion of faces detected was lower, classification was also less sensitive.

2.2.3. Human-human and human-machine inter-rater reliability

The gold standard for success of the automated system is not perfect agreement with a single human coder, but agreement at the

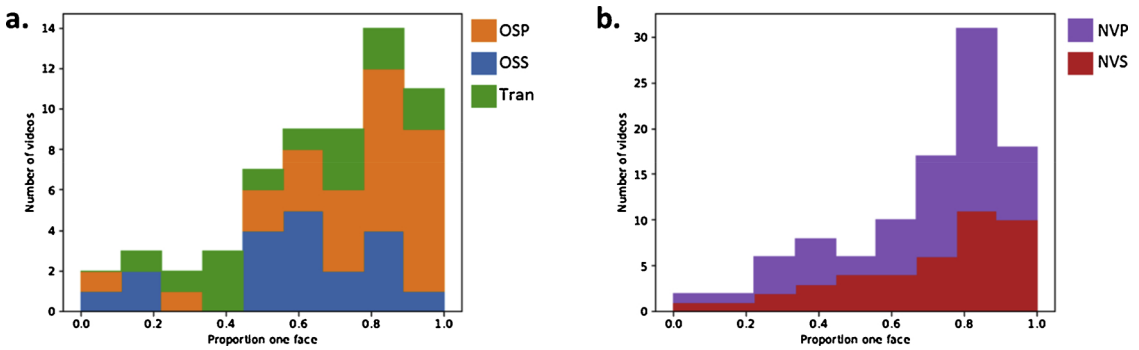


Fig. 1. Stacked histograms of the proportion of frames coded by human raters in which Rekognition detected one and only one infant face (PropOneFace), one value per video. (a) Experiment 1 (looking time). (b) Experiment 2 (preferential looking). OSP = OneShot publicity, OSS = OneShot Scientific, Tran = Tran et al. (2017), NVP = NovelVerbs publicity, NVS = NovelVerbs Scientific. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

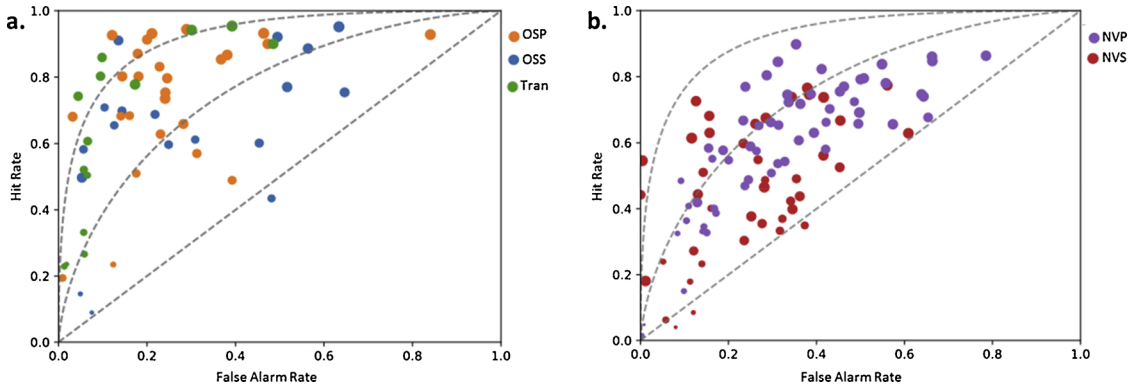


Fig. 2. Receiver operating characteristic (ROC) curves for (a) looking time and (b) preferential looking studies. ROC curves provide a visual representation of d-prime by plotting hit rate (hits divided by hits + misses) versus false alarm rate (false alarms divided by false alarms + correct rejections). The dashed grey curves show the sensitivities corresponding to d-prime = 0 (straight diagonal; indicates chance), d-prime = 1 and 2. The size of each dot represents the proportion of time that Rekognition detected exactly one infant face; bigger dots indicate greater PropOneFace values, which means a greater number of frames available to the classifier for training and rating (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

level that two human coders can generally achieve. We calculated Cohen's kappa, a measure of agreement that is suited to binary variables and is corrected for expected chance agreement, to quantify both human-classifier and human-human reliability for all pairs available. Note that the Tran data had a single rater only, precluding human-human kappa calculations; otherwise, average human-human and human-classifier kappa across subjects are reported in Table 3. For Experiment 1, although the average human-classifier kappa was significantly lower than the average human-human kappa (Table 3), there is some promise, in that according to Landis and Koch (1977) guidelines, the agreement is in the moderate range (0.41-0.60) for two experiments and fair (0.21-0.40) for another. Furthermore, the best human-classifier agreement (~0.75; bottom panel, Fig. 3a) was as strong as the weakest human-human agreement (top panel, Fig. 3a). For Experiment 2, the agreement is slight (0-0.20).

2.2.4. What factors influenced classification accuracy?

We considered potential predictors of two steps of the automated coding process: the ability to detect the infant's face using Rekognition (indexed by PropOneFace) and the ability to correctly classify gaze as looking or not looking based on Rekognition output (indexed by d-prime). We were primarily interested in the independent value of each predictor, rather than its incremental value in light of others, as each represents a separate variable to intervene on to improve either data collection or automated detection; therefore, we computed pairwise correlations rather than a regression. Correlations between the automated variables confidence, brightness, and sharpness were calculated for PropOneFace and d-prime (Table 4). PropOneFace was significantly and robustly correlated with confidence; there was a borderline correlation that did not survive correction for multiple comparisons with sharpness, and no correlation with brightness. That is, Rekognition's own confidence estimate was a reliable predictor of whether it successfully detected one infant face as expected. The higher-level index of classification sensitivity, d-prime, was not significantly correlated with brightness or sharpness; there was a borderline significant correlation with confidence which did not survive correction for multiple comparisons. None of the eight manually-coded video and child characteristics were significantly correlated with either d-prime (Fig. 4a) or PropOneFace (Fig. 5a) after correction for multiple comparisons; correlations are shown in Table 5.

Finally, as a check that the automated variables reflected the same characteristics we coded manually, we evaluated relationships between the manual factors 'amount of light' and 'resolution' with the automatically generated 'Brightness' and 'Sharpness' respectively. There were significant correlations between Brightness and manual 'amount of light' ($r = 0.52$, $p < 0.001$).

3. Experiment 2: detecting looks to left vs. right of the screen

In a preferential looking paradigm, two or more stimuli are presented, and total looking time to each is measured. It is a common developmental research paradigm, used to investigate everything from language to morality (see Golinkoff, Ma, Song, & Hirsh-Pasek,

Table 3
Average kappa across subjects (standard deviation in brackets). There is no Human-Human kappa statistics for Tran data as they were rated by a single human rater.

	Looking Time			Preferential Looking	
	OSP	OSS	Tran	NVP	NVS
Human-Human	0.91 (.09)	0.88 (.12)	n/a	0.79 (.14)	0.78 (0.12)
Human-Classifier	0.41 (.22)	0.30 (21)	0.46 (.21)	0.12 (.09)	0.11 (.11)

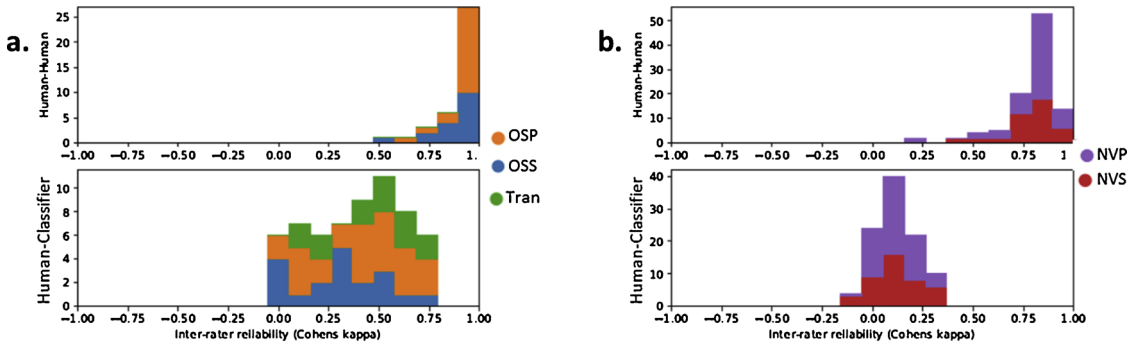


Fig. 3. Stacked histograms of kappa statistics quantifying inter-rater agreement on looking behavior coding. (a) Experiment 1 (looking time), (b) Experiment 2 (preferential looking). Top panel shows human-human inter-rater reliability; bottom panel illustrates human-classifier inter-rater reliability. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Table 4
Correlations comparing the automated Rekognition ratings "Confidence", "Brightness", and "Sharpness" and the measures "Dprime" and "PropOneFace", for Experiment 1 (looking time) and Experiment 2 (preferential looking). * indicates significance at $p < 0.05$ following by-experiment Bonferroni correction (raw $p < 0.017$).

	Exp 1 – Looking Time				Exp 2 – Preferential Looking			
	Dprime		PropOneFace		Dprime		PropOneFace	
	R	*P*	*r*	*p*	*r*	*p*	*r*	*p*
Confidence	.26	.04	.78	< .0005*	.20	.05	.70	< .0005*
Brightness	.15	.25	.15	.25	.03	.78	.27	.006*
Sharpness	−.04	.78	.28	.03	−.01	.95	.30	.002*

2013 and Tafreshi, Thompson, & Racine, 2014). In the NovelVerbs study, one video was shown on the left and one on the right of the screen. We evaluated automated scoring of whether the infant was looking "left" or right", which is more challenging than the looking time paradigm, as it requires discrimination of gaze positions within a smaller range of angles.

3.1. Methods

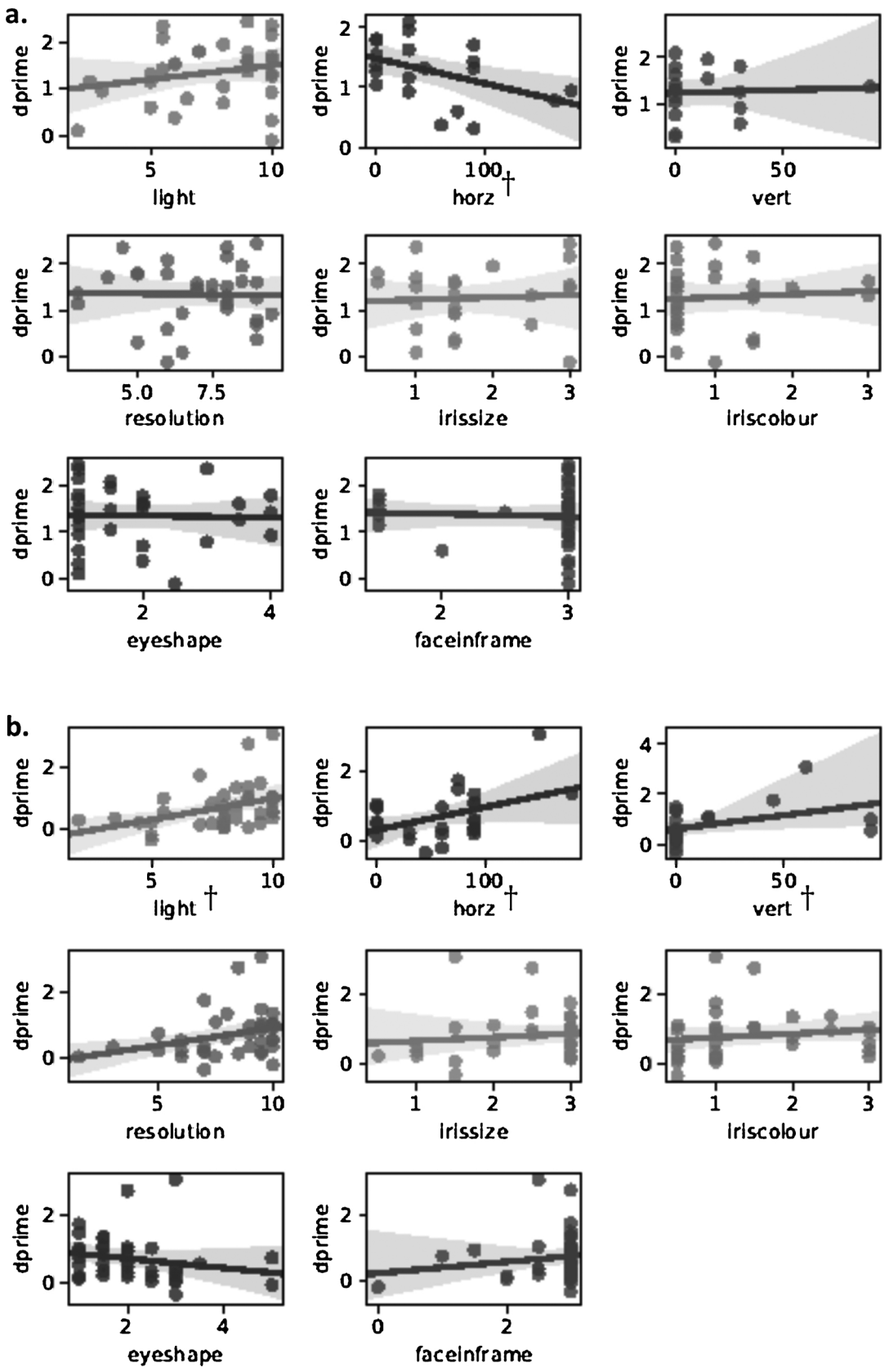
3.1.1. Video recordings

We used 100 pre-existing videos from Lookit's NovelVerbs study, from both the publicity (NVP; n = 42) and scientific level (NVS; n = 58) data.[1] Children between the ages of 24 to 36 months sat on their parents' laps to complete this looking-while-listening task. The session included three preferential looking segments: children were asked to find a familiar verb presented on the left and a familiar verb presented on the right, then asked to find a novel verb while potential target actions were shown on the left and right (Scott et al., 2017). Raters had previously annotated whether the child was looking left, looking right, or looking away from the camera, which was used to score automatic performance. The total number of minutes per video tagged by the manual raters ranged from 2 to 8 min (Table 1).

3.1.2. Face detection, pre-processing and classification

All 100 video samples were pre-processed as in Experiment 1. Again, to allow better generalization across child positioning relative to the webcam and screen, orientation and position features were zero-centered per child. This is expected to make coding of preferential looking particularly difficult, as it will effectively "subtract out" some genuine differences between children in overall preferences. However, we expect that looks *more* towards the right within a child will still be more likely to represent looking to the right, and to be coded as looking to the right, and vice versa. Although complete preferential looking coding requires determining not just whether the child is looking left or right but whether he or she is looking at all, for an initial proof of principle given these limitations we began with the simplified problem of distinguishing left from right within frames where a human coder had already coded the frame as 'left' or 'right' looking. The same classification method as in Experiment 1 was used, except that the possible classifications were left and right rather than looking and not looking. If a face was not detected, we defaulted to "left".

[1] There were 14 instances of videos that had been lightened to increase visibility of the child's eyes for the human raters, and as in the looking time study, we used the original rather than the lightened videos.



(caption on next page)

Fig. 4. Pairwise correlations between classifier sensitivity (d-prime) and manual ratings of video characteristics. (a) Experiment 1 (top panel; looking time), (b) Experiment 2 (bottom panel; preferential looking). Manually coded variables, described in Table S1, are amount of light (light), horizontal (horz), vertical (vert), resolution, iris size (irissize), iris colour (iriscolour), eye shape (eyeshape), and face in frame (faceinframe). † indicates p value ≤ .05; *indicates $p < 0.05$ following by-experiment Bonferroni correction (raw $p < 0.006$).

3.2. Results

As in Experiment 1, PropOneFace was calculated (Fig. 1b) and the performance of the classifier on the features from Rekognition was compared to previously acquired manual ratings.

3.2.1. Were infant faces automatically detected?

The range and average for PropOneFace are in Table 1. As in Experiment 1, Rekognition detected a single infant face in about 70% of video frames on average, although success varied widely across videos. There was no difference between the subsamples NVS and NVP, $t(98) = .159$, $p = .874$, two-tailed.

3.2.2. Could left vs. right be classified?

Again, d-prime was calculated to assess sensitivity of the automated tagging, with higher d-prime indicating better sensitivity (Table 1). The sensitivity (d-prime) on this task was lower than when coding looking time in Experiment 1, $t(105) = 6.813$, $p < .001$. Although performance was above chance overall, performance on many individual videos was close to chance (points close to the diagonal on the ROC plot, Fig. 2b). The sensitivity metric, d-prime, for NVP and NVS did not differ, $t(54) = .688$, $p = .495$, two-tailed, with an overall d-prime of M = .76, SD = .52. There was a significant correlation between PropOneFace and d-prime ($r = .29$, $p = .003$, two-tailed) for Experiment 2 overall, indicating that failure to code gaze direction correctly was at least in some cases related to failure to detect the infant face.

Kappa statistics for human-human reliability (top panel) and human-Rekognition reliability (bottom panel) are shown in Fig. 3b. Human-classifier agreement was weak with most Kappas < 0.3.

3.2.3. Were there factors that influenced face detection and classification accuracy?

Table 4 shows correlation coefficients and p-values for relationships between the success metrics PropOneFace and d-prime and the three automatic variables confidence, brightness, and sharpness. PropOneFace was robustly correlated with all three automatic variables confidence, brightness, and sharpness. However, as in Experiment 1, correlations with d-prime were weaker. There was a positive relationship between d-prime and confidence, although it did not survive correction for multiple comparisons, and no significant relationship with either brightness or sharpness.

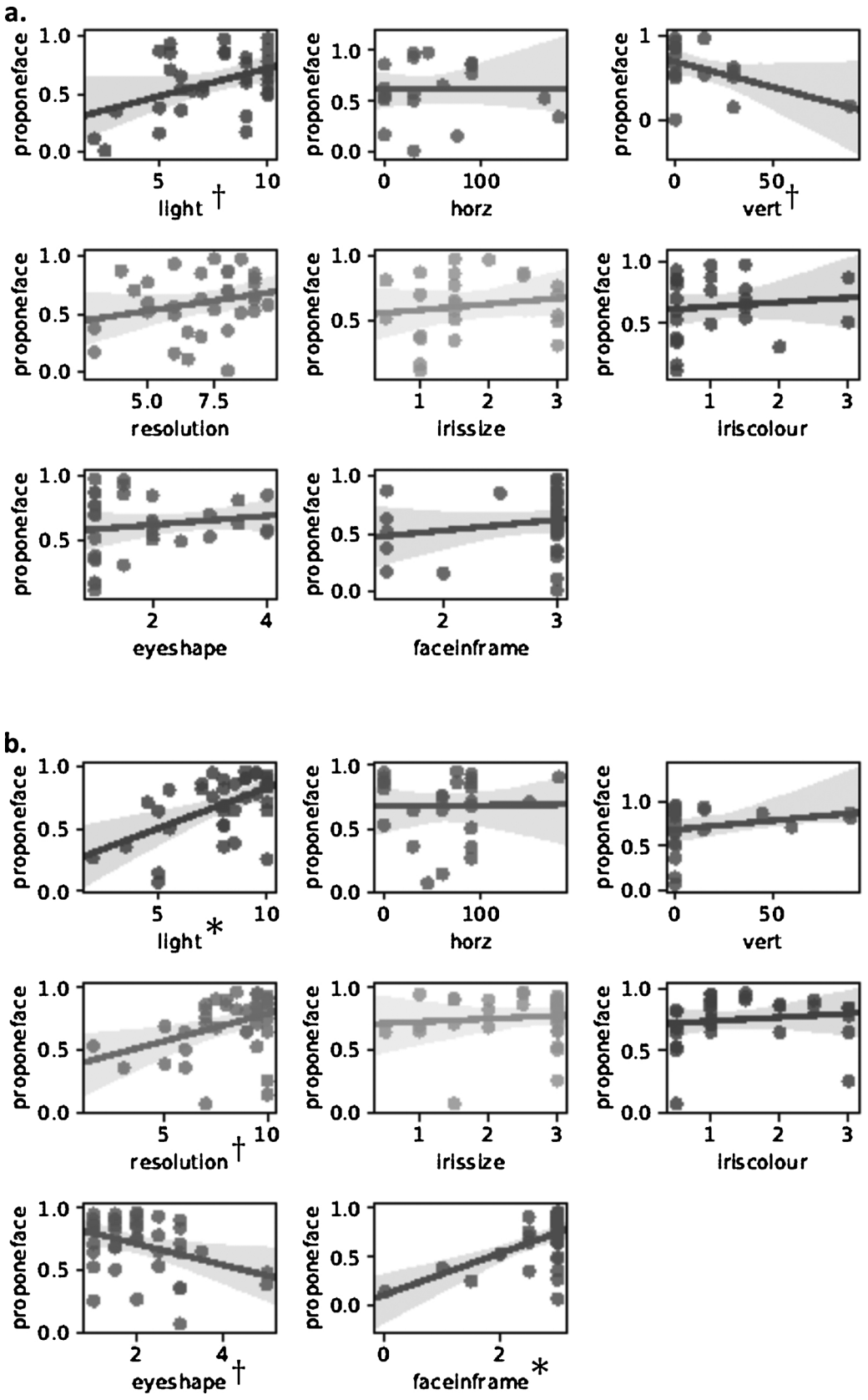
The far-right columns of Table 5 contain the correlations for each of the manually-coded video characteristics with d-prime (Fig. 4b) and with PropOneFace (Fig. 5b) in Experiment 2. PropOneFace was significantly correlated with amount of light and face in frame, while resolution and eye-shape failed to survive correction for multiple comparisons (both $ps = .02$). Again, d-prime was not significantly correlated with any of the manually-coded characteristics following correction for multiple comparisons, although there were some borderline correlations (amount of light, horizontal, vertical, all $ps \leq .05$). As expected, manually-rated brightness correlated with automatically-rated brightness ($r = 0.42$, $p = 0.006$).

4. Discussion

A commercially-available face analysis tool, combined with a classifier, was in many ways successful in automatically scoring videos of infants acquired online. The resulting cheap, automated pipeline for acquisition and analysis has the potential to improve replicability in developmental science by boosting sample sizes and allows for transparent sharing of code that includes recruitment. We also acquired measures of video quality with the intention of providing ways to optimize future data acquisition and improve automatic scoring. It should be noted that the data used here were acquired with the intention of performing manual rating, and many possible improvements are practical and that the classifier was evaluated on videos collected online, but would also be applicable to videos that were collected in-person.

Automated scoring was reasonably accurate for coding looking towards or away from the screen. In Experiment 1, d-prime sensitivity was 1.00–1.75 and the highest kappas obtained were around 0.75, which falls at the high end of 'substantial' in terms of agreeing with human acquired ratings (Landis & Koch, 1977). While this is promising, the majority of machine-human agreement values were lower than the human-human agreement (typically > 0.9). Less promising, our findings for preferential looking coding (left vs. right) show substantial room for improvement. Although above chance, sensitivity was quite low, with d-prime values < 0.8, and the machine-human agreement was poor (kappa generally < 0.3).

There were two things that had to happen in order for the classifier to do its job well. In the first stage, one and only one infant face had to be detected in the still frames that were used for each video (PropOneFace). These frames were fed into the classifier (stage 2), which had to then classify the infant face in each frame as looking/not looking (Exp 1) or looking left vs. right (Exp 2). If Rekognition was only successful at recognising one and only one infant face in very "pristine" frames, then we would expect all of the PropOneFace frames fed forward to be pristine and thus easily coded by our classifier. This would have been indicated by very high d-prime values for the entire range of PropOneFace scores, which was not what we found. Rather, the positive correlations in both



(caption on next page)

Fig. 5. Pairwise correlations between Rekognition measure PropOneFace and manual ratings of video characteristics. (a) Experiment 1 (looking time), (b) Experiment 2 (preferential looking). Manually coded variables, described in Table S1, are amount of light (light), horizontal (horz), vertical (vert), resolution, iris size (irissize), iris colour (iriscolour), eye shape (eyeshape), and face in frame (faceinframe). † indicates p value ≤ .05; *indicates $p < 0.05$ following by-experiment Bonferroni correction (raw $p < 0.006$).

Table 5
Correlations comparing the manual ratings "Amount of light", "Horizontal", "Vertical", "Resolution", "Iris size", "Iris colour", "Eye shape", and "Face in frame" to the Rekognition measures "Dprime" and "PropOneFace", for Experiment 1 (looking time) and Experiment 2 (preferential looking). * indicates significance at $p < 0.05$ following by-experiment Bonferroni correction (raw $p < 0.006$).

	Exp 1 – Looking Time				Exp 2 – Preferential Looking			
	Dprime		PropOneFace		Dprime		PropOneFace	
	r	*p*	*r*	*P*	*r*	*p*	*r*	*p*
Amount of light	.23	.21	.45	.009	.40	.01	.55	< .001*
Horizontal	.44	.05	.00	.998	.43	.03	.01	.96
Vertical	.06	.83	−.48	.04	.42	.05	.23	.30
Resolution	−.02	.92	.25	.17	.31	.06	.39	.02
Iris size	.06	.78	.15	.48	.11	.53	.10	.59
Iris colour	.07	.73	.11	.60	.14	.45	.14	.44
Eye shape	−.03	.88	.15	.41	−.21	.20	−.38	.02
Face in frame	−.05	.77	.20	.26	.15	.34	.55	< .001*

experiments indicated that when there was difficulty detecting one and only one infant face (i.e., low PropOneFace), there was also difficulty with classification sensitivity (i.e., low d-prime). Somewhat similarly, the positive correlation also reflects that more information improved our classifier accuracy, thus supporting the notion that the classifier was driving the relationship, because if the classifier was always highly accurate (or inaccurate), there would be no correlation. Altogether, the findings suggest that in the future, improved PropOneFace will most likely improve classifier performance. In addition, the findings may also indicate that certain characteristics of the child and/or video might influence both stages.

We have preliminary data investigating different video and child characteristics that may have influenced Rekognition's ability to locate the infant face and/or the classifier's ability to code for looking/not looking and for looking left/right. In both experiments, brighter video predicted a better ability to detect one and only one infant face (i.e., PropOneFace). In Experiment 2, PropOneFace was positively correlated with higher video resolution, a greater amount of the child's face in the frame, and larger or more open eyes (although correlations did not survive correction for multiple comparisons). However, there were fewer correlations with d-prime, and these were weaker, which makes sense in light of the greater complexity of classifying gaze of eyes as opposed to presence of face and the fact that coding ability depends on details of view and setup even for human coders. Regardless, both video brightness and a light source coming from in front of the subject weakly predicted sensitivity. Altogether, the data here suggest that different video and child characteristics influence PropOneFace and d-prime, but that well-lit environments and avoiding backlighting during data collection should be evaluated as ways to improve potential of both stages of automated coding.

A benefit of choosing Amazon Rekognition was that it estimates the age of the faces detected, which was successful in identifying infant faces for all 160 of our videos. We used a liberal age criterion of less than 10 years, which worked well for our data in which an infant sat on their parent's lap. However, the same criterion may be too liberal to identify the participant in paradigms that include older or multiple children.

We obtained this weak but promising classification performance on videos in which no a priori consideration had been given to ways to maximize success of automated tagging. Results may have been more positive if the classifier had been evaluated on higher quality videos that were collected in the lab. Regardless, it is apparent from the current study that there are potential ways to improve the quality of videos collected online that would, in turn, enhance Rekognition's automated tagging sensitivity. One potential solution would be to have thresholds of acceptability that could be monitored in real time and fed back to parents. For example, both human and Rekognition ratings of brightness correlated with Rekognition's ability to consistently detect an infant face throughout a video (i.e., PropOneFace). Future studies could begin with a five second video clip immediately rated for brightness by Rekognition and that must pass a certain threshold in order for the study to proceed. Additionally, information regarding how best to achieve ideal lighting and positioning (e.g., position of lighting, artificial vs. natural light, bounding box for the child's face in the webcam stream) could be offered as tips to the parents as they set up for the study.

We emphasize that obtaining *any* ability to automatically score videos is promising for future work, given the simplicity of this approach: in particular, our classifier simply used linear combinations of feature positions in the video, without construction of a 3D model of the setup; and generalization was required across children with *no* within-child training data used, despite known differences in children's positioning relative to the screen – meaning that the same position of features is not expected to represent the same gaze target, and that a given movement of the features is not expected to represent the same distance between targets.

Future work has the potential to improve on classification algorithms in several directions. Approaches may be based on existing open-source tools such as Web Gazer (https://webgazer.cs.brown.edu/), other computational methods (Harari, Gao, Kanwisher, Tenenbaum, & Ullman, 2016) or algorithms designed for eye tracking using specialized hardware (such as Tobii; https://www.

tobiipro.com/). First, given that face detection is a necessary intermediate step for gaze coding, minimal human coding – for instance, providing bounding boxes for the child's face in several frames, or selecting the participant's face from those detected – could be used to support automated approaches, while still dramatically reducing the amount of human effort required. Second, both face detection and coding might be improved by making use of the face that data is collected as video, rather than individual still frames. Human coders rely heavily, for instance, on the assumption that the gaze target is heavily autocorrelated in time, and on phenomena such as the child's ability to move his or her head without shifting the gaze target.

Third, there are a variety of potential approaches that might enable more successful extraction of gaze direction based on Rekognition or other feature detection output. To get more accurate boundaries between gaze categories for each child, given that as noted children's gaze directions are unlikely to each be centered on the screen center, we could apply non-linear and/or model-based approaches; we could make use of either calibration or a small amount of human-coded data within children; we could select a classifier trained on data where children were positioned similarly; and/or, as human coders do, we could use the distribution of estimated gaze target positions to infer category boundaries, for instance expecting two "clusters" of gaze positions when conducting a preferential looking study.

Robust algorithmic approaches to coding infant looking behavior from video will find many potential applications. Freely-available well-validated algorithms for automated coding have the potential to improve replicability of analysis across labs: while most researchers report interrater reliability for a subset of their data using two raters from the same lab, to our knowledge no data is available regarding systematic variation across labs in gaze coding, even prior to processing to extract dependent measures such as the looking time until the first lookaway. Variation in other subjective judgments such as fussiness (Slaughter & Suddendorf, 2007), however, suggests potential for inter-lab differences. Collaborative efforts such as the ManyBabies studies (Frank et al., 2017) and video sharing on Databrary (Simon, Gordon, Steiger, & Gilmore, 2015) may shed light on the degree of systematic differences. In addition to enabling fully-automated online data collection and analysis pipelines as described here, automated coding would eliminate a major bottleneck in lab-based research and enable researchers to design and use infant-contingent experimental protocols, for instance training or habituation paradigms, online or in the lab without relying on an experimenter's live coding.

5. Conclusions

This study provides a proof-of-principle and sets a baseline for automated scoring of video collected online, which extends to videos collected in the lab. Amazon Rekognition is a promising technology for the automated rating of video using infant looking-time paradigms. Related approaches may in the future remove the considerable burden of human rating to facilitate more replicable developmental science.

Declarations of interest

None.

Acknowledgements

B.C. is supported by funding from the charity RESPECT and the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007–2013) under REA grant agreement no. PCOFUND-GA-2013-608728.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.infbeh.2018.11.004.

References

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., ... Zuni, K. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), https://doi.org/10.1126/science.aac4716.

Amazon (2017). *Amazon rekognition developer guide.* Retrieved April 18, 2018, fromhttps://docs.aws.amazon.com/rekognition/latest/dg/what-is.html.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3–5. https://doi.org/10.1177/1745691610393980.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., ... Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376. https://doi.org/10.1038/nrn3475.

Cooper, R. P. (1990). Preference for Infant-Directed Speech in the First Month after Birth Author (s): Robin Panneton Cooper and Richard N. Aslin Published by: Wiley on behalf of the Society for Research in Child Development Stable http://www.jstor.org/stable/1130766, 61(5), 1584–1595.

Ebert, J. F., Huibers, L., Christensen, B., & Christensen, M. B. (2018). Paper- or web-based questionnaire invitations as a method for data collection: Cross-sectional comparative study of differences in response rate, completeness of data, and financial cost. *Journal of Medical Internet Research, 20*(1), e24. https://doi.org/10.2196/jmir.8353.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy, 22*(4), 421–435. https://doi.org/10.1111/infa.12182.

Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: What have we learned? *Perspectives on Psychological Science, 8*(3), 316–339. https://doi.org/10.1177/1745691613484936.

Granello, D., & Wheaton, J. (2004). Online data collection: Strategies for research. *Journal of Counseling & Development, 82*(4), 387–393. https://doi.org/10.1002/j.1556-6678.2004.tb00325.x.

Hagedorn, J., Hailpern, J., & Karahalios, K. (2008). VCode and VData: Illustrating a new framework for supporting the video annotation workflow. *Proceedings of the Workshop on Advanced Visual Interfaces AVI,* 317–321. https://doi.org/10.1145/1385569.1385622.
Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*(7169), 557–559. https://doi.org/10.1038/nature06288.
Harari, D., Gao, T., Kanwisher, N., Tenenbaum, J., & Ullman, S. (2016). Measuring and modeling the perception of natural and unconstrained gaze in humans and machines. *arXiv.org*(59), arXiv:1611.09819. Retrieved from http://arxiv.org/abs/1611.09819v1.
Kipp, M. (2014). ANVIL: A universal video research tool. In J. Durand, U. Gut, & G. Kristofferson (Eds.). *Handbook of Corpus phonology* (pp. 420–436). Oxford University Press.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310.
Lefever, S., Dal, M., & Matthíasdóttir, Á. (2007). Online data collection in academic research: Advantages and limitations. *British Journal of Educational Technology, 38*(4), 574–582. https://doi.org/10.1111/j.1467-8535.2006.00638.x.
Lonsdale, C., Hodge, K., & Rose, E. A. (2006). Pixels vs. paper: Comparing online and traditional survey methods in sport psychology. *Journal of Sport & Exercise Psychology, 28*(1), 100–108.
Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101–B111. https://doi.org/10.1016/S0010-0277(01)00157-3.
McDuff, D., El Kaliouby, R., & Picard, R. W. (2015). Crowdsourcing facial responses to online videos: Extended abstract. *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015, 3*(4), 512–518. https://doi.org/10.1109/ACII.2015.7344618.
Montague, D. P. F., & Walker-Andrews, A. S. (2001). Peekaboo: A new look at infants' perception of emotion expressions. Developmental Psychology. Montague, Diane P.F.: Long Island U, Dept of Psychology, 1 University Plaza, Brooklyn, NY, US, 11201, diane.montague@liu.edu: American Psychological Association. https://doi.org/10.1037/0012-1649.37.6.826.
Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie Du Sert, N., ... Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021.
Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making, 5*(5), 411–419. https://doi.org/10.2139/ssrn.1626226.
Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*(6), 531–536. https://doi.org/10.1177/1745691612463401.
Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology, 299*, 172–179. https://doi.org/10.1016/j.jtbi.2011.03.004.
Resnik, D. B. (2006). Openness versus secrecy in scientific research. *Episteme, 2*(3), 135–147. https://doi.org/10.3366/epi.2005.2.3.135.
Riva, G., Teruzzi, T., & Anolli, L. (2003). The use of the internet in psychological research: Comparison of online and offline questionnaires. *CyberPsychology & Behavior, 6*(1), 73–80. https://doi.org/10.1089/109493103321167983.
Scott, K., Chu, J., & Schulz, L. (2017). Lookit (Part 2): Assessing the viability of online developmental research, results from three case studies. *Open Mind: Discoveries in Cognitive Science, 1*(1), 15–29. https://doi.org/10.1162/opmi_a_00001.
Scott, K., & Schulz, L. (2017). Lookit (Part 1): A new online platform for developmental research. *Open Mind: Discoveries in Cognitive Science, 1*(1), 1–14. https://doi.org/10.1162/opmi_a_00002.
Semmelmann, K., Hönekopp, A., & Weigelt, S. (2017). Looking tasks online : Utilizing webcams to collect video data from home. *Frontiers in Psychology, 8*(September), 1–11. https://doi.org/10.3389/fpsyg.2017.01582.
Simon, D. A., Gordon, A. S., Steiger, L., & Gilmore, R. O. (2015). Databrary: Enabling sharing and reuse of research video. *Proceedings of the 15th ACM/IEEE-CE on joint conference on digital libraries - JCDL' 15,* 279–280. https://doi.org/10.1145/2756406.2756951.
Slaughter, V., & Suddendorf, T. (2007). Participant loss due to "fussiness" in infant visual paradigms: A review of the last 20 years. *Infant Behavior & Development, 30*, 505–514. https://doi.org/10.1016/j.infbeh.2006.12.006.
Tafreshi, D., Thompson, J. J., & Racine, T. P. (2014). An analysis of the conceptual foundations of the infant preferential looking paradigm. *Human Development, 57*(4), 222–240.
Tran, M., Cabral, L., Patel, R., & Cusack, R. (2017). Online recruitment and testing of infants with Mechanical Turk. *Journal of Experimental Child Psychology, 156*, 168–178. https://doi.org/10.1016/j.jecp.2016.12.003.
Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69*(1), 1–34. https://doi.org/10.1016/S0010-0277(98)00058-4.