A Quantitative Comparison of Manual vs. Automated Facial Coding Using Real Life Observations of Fathers

Romana Burgess[1,2(✉)], Iryna Culpin[2,3], Helen Bould[2,4,5], Rebecca Pearson[2,3], and Ian Nabney[1]

[1] Digital Health Engineering Group, Faculty of Engineering, Merchant Venturers Building, University of Bristol, Bristol, UK
romana.burgess@bristol.ac.uk
[2] Centre for Academic Mental Health, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
[3] Department of Psychology, Manchester Metropolitan University, Manchester, UK
[4] Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, UK
[5] Gloucestershire Health and Care NHS Foundation Trust, Gloucester, UK

Abstract. This work explores the application of an automated facial recognition software "FaceReader" [1] to videos of fathers ($n = 36$), taken using headcams worn by their infants during interactions in the home. We evaluate the use of FaceReader as an alternative method to manual coding – which is both time and labour intensive – and advance understanding of the usability of this software in naturalistic interactions. Using video data taken from the Avon Longitudinal Study of Parents and Children (ALSPAC), we first manually coded fathers' facial expressions according to an existing coding scheme, and then processed the videos using FaceReader. We used contingency tables and multivariate logistic regression models to compare the manual and automated outputs. Our results indicated low levels of facial recognition by FaceReader in naturalistic interactions (approximately 25.17% compared to manual coding), and we discuss potential causes for this (e.g., problems with lighting, the headcams themselves, and speed of infant movement). However, our logistic regression models showed that when the face was found, FaceReader predicted manually coded expressions with a mean accuracy of $M = 0.84$ (*range* = 0.67–0.94), sensitivity of $M = 0.64$ (*range* = 0.27–0.97), and specificity of $M = 0.81$ (*range* = 0.51–0.97).

Keywords: Automated facial coding · FaceReader · ALSPAC

1 Introduction

Manual coding is a comprehensive method for capturing facial expressions from observational data, and is easily adaptable to a range of contexts and scenarios. However, manual coding is both time and labour intensive, and can be biased by the experiences of the human coder (for example, the amount of time spent coding, or the previous

A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 379–396, 2023.
https://doi.org/10.1007/978-3-031-34586-9_25

expression coded). These disadvantages may be addressed by automated facial coding, which offers rapid and detailed decomposition of facial expressions. Automated facial coding could potentially cut down on time spent manually coding, and also reduce any biases. This would allow for more data to be processed, potentially more accurately.

One such automated facial coding software is the Noldus FaceReader [1] which was first proposed by Den Uyl & Van Kuilenburg [2], who described the process involved in finding, modelling, and classifying a face. The software uses deep learning and neural networks to classify faces into one of eight facial expressions: Happy, Sad, Angry, Scared, Surprised, Disgusted, Neutral and Contempt.

A validation study was performed on the software [3], comparing human and FaceReader facial classification for two publicly available, objective datasets of human expressions. On average, FaceReader correctly identified 89% of expressions, whereas human coders correctly identified 85%. This work also identified variation in accuracy across different expressions, e.g., FaceReader classified Happy with 96% accuracy, but classified Angry with 76% accuracy. Another validation was performed on a later version of the software [4], finding that - on average across all expressions - 80% of expressions were correctly classified. Other reported rates of performance for the software are 89% [2] and 87% [5]. One study observed gender-based differences [6], noting that FaceReader better identified Surprised and Scared emotions in males, and better identified Disgusted and Sad emotions in females.

Previous works have compared the performance of FaceReader to human coding. For example, one study investigated the expressions of students taking a test, measuring the agreement between two human coders and FaceReader [6]. This work found that the humans and the software agreed strongly for Neutral and Happy expressions, agreed often for Sad, Scared, and Surprised, and did not agree often for Disgust and Angry. A similar finding was reflected in [5], who found that agreement was highest between FaceReader and manual coders for Neutral and Happy, and lowest for Angry and Disgust. There have been many similar applications of the FaceReader software across different domains, including: evaluating human reactions to complex web-based tasks [7, 8], measuring implicit and explicit expressions during orange juice tasting sessions [9], and evaluating spontaneous expressions vs. posed expressions [10].

However, previous applications of FaceReader have commonly used videos filmed in controlled environments with good lighting, and a homogenous background with no people or objects [9–11]. To be useful for many real-life applications, it is vital that expression recognition is effective for naturalistic, uncontrolled environments, such as within the home. While some have implemented other methods to recognise faces for "in-the-wild" videos [12], we have found very few studies where FaceReader has been applied to these kind of observations. We identified one example [13] which used FaceReader to analyse a dataset containing movie clips of actors, and found that the software could not accurately classify any expression.

Many studies have also used videos that were recorded using built-in laptop webcams, providing a direct view of the participant's face. Naturalistic interactions – for example involving multiple people or different body positions – may not be well captured by a webcam, or any kind of stationary camera. Wearable headcams provide an ideal solution for capturing facial expressions in a naturalistic setting, and may enable

more ecologically valid interactions, e.g., by containing less socially desirable facial expressions than in a controlled observation [14]. We have not identified any studies applying FaceReader analysis, or any other automated facial coding, to videos taken using wearable headcams during natural interactions.

FaceReader has rarely been explored in a parent-infant context. One study used the software to analyse the intensity of mothers' Happy expressions during exposure to images of infants [15], and another carried out five separate tests using FaceReader in a parent-infant context [16]. These tests analysed facial expressions across different scenarios, including mother-infant interactions, infant-infant interactions, and interactions in infants with developmental disorders. All except one test used observations carried out in naturalistic settings with uncontrolled lighting and a handheld video camera (one test used a laboratory setting with controlled lighting). Videos were excluded from analysis if the participants head rotation was greater than 45 degrees from the camera. The authors highlighted that assessing facial expressions in infant interactions is vital to understand the "emotional sphere" of the child, with the goal of identifying and addressing less optimal emotional responses (e.g., smiling at an infant cry) [16]. While these studies both contributed to understanding maternal expressions, we did not find any FaceReader literature specifically studying fathers' facial expressions.

Father-infant interactions have been studied much less than those between mothers and infants. This may be in part because fathers have traditionally been less involved in childcare, although modern fatherhood roles are evolving to include more social, emotional, and physical childcare than ever before [17]. Yet, studies of fathers are valuable, as father-infant interactions have been found to contribute to infant language and cognitive development [18] and to be predictive of behavioural problems [19]. Infants begin to develop emotional coordination – learning to discriminate and respond to emotional expressions, vital for social function [20] – from as early as 4 months old [21]. Whilst much is known about how emotional coordination occurs in mother-infant interactions, less is understood about fathers [22]. There are likely to be differences in the communicative mechanisms, and by observing, quantifying, and analysing father and infant facial expressions during an interaction, we can begin to understand these differences.

The aims of our work were: (1) to evaluate the performance of FaceReader on videos of naturalistic father-infant interactions captured using a wearable headcam, and (2) to evaluate the relationship between the automated and human coding of facial expressions. To address these aims, we coded 36 videos of fathers engaging in free play or feeding interactions with infants, both manually and using FaceReader. We then used contingency analysis and logistic regression classification models to compare the two relative outputs. Through this work, we provide new information regarding the use of FaceReader to process paternal facial expressions in a naturalistic setting.

2 Methodology

2.1 Data

We used data taken from the Avon Longitudinal Study of Parents and Children (ALSPAC). The study website (http://www.bristol.ac.uk/alspac/researchers/our-data/) contains details of all ALSPAC data that are available through a fully searchable data

dictionary and variable search tool. ALSPAC data are collected and managed using Research Electronic Data Capture (REDCap) electronic data capture tools hosted at the University of Bristol [23]; REDCap is a secure web-based platform designed to support data capture for research studies.

Full ALSPAC cohort demographics and recruitment details have been provided previously elsewhere [24–27]. In brief, ALSPAC is an ongoing longitudinal, population-based study based in Bristol, UK. The original cohort was recruited via 14 541 pregnancies with expected delivery dates between 1 April 1991 and 31 December 1992; this original cohort is referred to as generation 0, or ALSPAC-G0. The children born in 1992 to the ALSPAC-G0 cohort are referred to as generation 1, or ALSPAC-G1. And finally, the children born to the ALSPAC-G1 cohort in recent years are referred to as generation 2, or ALSPAC-G2. Our work comprises videos of fathers from ALSPAC-G1 (whose infants are in ALSPAC-G2). The fathers in this work had a mean age of 31.31 years ($SD = 5.45$), and their infants had a mean age of 32.62 weeks ($SD = 5.85$). Eight infants were male, and five infants were female.

Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time.

The videos were taken using headcams worn by infants during father-infant interactions within the home. Previous work has found first-person headcams to be reliable for capturing behaviours during parent-infant interactions [14]. Fathers were recruited through a father-specific research clinic, inviting dads to take part in several assessments when their child turned six months old. One of these assessments was the headcam study, for which there were no specific selection criteria. Table 1 outlines the video contributions from each father included in our work.

Table 1. Detail about the videos provided by each father within the dataset.

ID	# Videos	Total length (sec)	Interaction types (# of videos)
F1	1	789	Feeding (1)
F2	2	736	Feeding (1), Free play (1)
F3	2	1141	Feeding (1), Free play (1)
F4	6	2776	Feeding (4), Free play (2)
F5	3	1656	Feeding (3)
F6	1	430	Free play (1)
F7	6	2325	Feeding (3), Free play (2), Combination (1)
F8	4	1396	Feeding (2), Free play (1), Combination (1)
F9	2	782	Feeding (1), Free play (1)
F10	1	916	Feeding (1)

(*continued*)

Table 1. (*continued*)

ID	# Videos	Total length (sec)	Interaction types (# of videos)
F11	3	2136	Feeding (2), Free play (1)
F12	2	1263	Feeding (2)
F13	3	1022	Feeding (1), Free play (1)
Total	**36**	**17 368**	**Feeding (24), Free play (10), Combination (2)**

We used 36 videos in total, collected between 2019–2020, including: 24 feeding interactions, 10 free play interactions, and 2 videos which were a combination of both feeding and free play. These videos come from 13 individual fathers, as many provided multiple separate videos (see Table 1).

As this work is not about individual differences in expressions, but rather the efficacy and accuracy of the software at capturing expressions, we found no reason to exclude second or third videos from a single participant. Also, as all videos were varied in length, it was possible that one father may include multiple videos roughly equating to the length of a single video from another father (e.g., F1 and F2).

2.2 Manual Coding

All videos were manually coded using Noldus Observer 15 software [28], at a temporal resolution of 1/5 s. The facial expressions were coded according to the MHINT coding scheme [29], which is freely available to access online. Specifically, this includes the following expressions: Smile, Positive, Neutral/Alert, Negative, Surprise, Mock Surprise, Woe Face, Disgust, None of the Above and Face not Visible. These expressions are exhaustive and mutually exclusive, meaning that every timestamp is allocated a unique, single expression.

Initial coding was performed by one researcher, and two additional researchers were recruited for double coding. Seven randomly selected videos were selected for double coding, with one researcher coding four videos, and one researcher coding three. This equated to 3906 s of double-coded data, or 22.38% of the total video data. All reliability analyses were conducted using the Observer XT 15.0 [28].

To measure inter-coder agreement, we used the *index of concordance*. This is calculated by the total agreement for a behaviour (i.e., the duration that an expression is coded as present/not present by both coders) divided by the total duration of the interaction. The index of concordance is expressed as a value between 0 (no agreement) and 1 (total agreement). Across all expressions, an index of concordance of 0.93 ($SD = 0.07$) was achieved with the first double coder, and 0.91 ($SD = 0.07$) was achieved with the second coder. Inter-coder results by facial expression are shown in Table 2. In these analyses, we excluded expressions that occurred for less than 1% of the total interaction duration (i.e., Woe face, Disgust, Surprise).

Table 2. Mean inter-coder reliability by facial expression (*SD*).

Neutral	Positive	Smile	Negative	Mock Surprise	None of the above	Face not visible
0.89 (0.09)	0.89 (0.09)	0.96 (0.03)	0.98 (0.00)	0.98 (0.02)	0.94 (0.05)	0.89 (0.08)

2.3 Automated Facial Coding Using FaceReader

All videos were also processed using Noldus FaceReader [1], a facial recognition software trained to classify eight facial expressions in adult humans: Happy, Sad, Neutral, Angry, Scared, Surprised, Disgusted, Contempt. To make this classification, deep learning algorithms are first used to find a face within an image, while varying for facial position and size. Eye tracking is also used to help identify the rotation of the face. Based on deep neural networks, an artificial face is then synthesised using around 500 key points on the face; these describe the position of features and muscles, as well as different textures (e.g., eyebrow presence).

Expression classification takes place using a neural network, trained to recognise facial expressions using over 20,000 manually annotated images of faces [30]. When presented with a face, the software fits a "mesh" over the face and its key points, then calculates the deviation of these points from their position relative to a "mean" face, in order to make a prediction of the expression. A more detailed explanation of the FaceReader software can be found elsewhere [31].

Once a face has been detected, FaceReader provides multiple detailed outputs describing the facial expression present within that frame. The software processes videos frame-by-frame, which in our case resulted in an output being provided for every 0.033 s. In this work, we use the FaceReader output *expression intensity*: a single value within the interval [0, 1] describing the strength of each of the eight expressions. An intensity close to 0 indicates the expression is not present, and an intensity close to 1 indicates an expression is very present. An intensity value is provided for each of the eight expressions simultaneously, with each value independent of one another (i.e., the values do not sum to one).

It should be noted that the FaceReader expressions do not directly match those within the manual coding scheme (e.g., "Mock Surprise" is a manual expression, but not a FaceReader one), however, this was not problematic for the purposes of our work. Table 3 shows an approximate mapping between the manual and FaceReader expressions.

Table 3. Approximate mapping between manual and FaceReader expressions.

Manual Expression(s)		FaceReader Expression(s)
Neutral/Alert	→	Neutral
Smile + Positive	→	Happy
Negative	→	Sad + Angry + Scared + Contempt
Surprise + Mock surprise	→	Surprised
Disgust	→	Disgusted
None of the Above + Woe face	→	*n/a*
Face not visible	→	Face not found

2.4 Data Analysis

Data Pre-processing. Before we started the data analysis, we first carried out some pre-processing. A flow diagram showing all pre-processing stages is provided in the Appendix. We started by removing all data where a second caregiver was present during the interaction. This typically happened when the mother came to bring food, to admire the headcam on the infant, or to walk past in the background. These data were removed because FaceReader would often mistakenly classify the facial expression of the second caregiver during these periods, rather than that of the father. This meant that amount of viable coded data reduced from 17,368 s to 15,420 s.

The expression intensities were also normalised during pre-processing. This was necessary because the manual coder can only choose one dominant facial expression at a time, so for consistency, we must assume that we cannot have multiple dominant expressions. By normalising the intensities, this helps to highlight the dominant expression within the FaceReader output.

Data Analysis Procedures. All analyses were carried out using Python 3.0 [32]. Our aims were twofold: (1) to evaluate the performance of FaceReader on videos of naturalistic father-infant interactions captured using a wearable headcam, and (2) to evaluate the relationship between the automated and human coding of facial expressions.

To address aim (1), we calculated the amount of time that a face was detected and classified by both the human coder and the FaceReader software. We also calculated the amount of time that a face was not detected by both the human and the software. These values are displayed in a contingency table (Table 4) in Sect. 3.1.

To address aim (2), we used multivariate binary logistic regression; a choice made due to the simplicity of the model and the ease of parameter interpretation. Logistic regression measures the probability of a data entry being classified as one of two mutually exclusive, exhaustive states (which we assign as either a 0 or a 1). In our work, this translates to the eight simultaneous, FaceReader expression intensities being classified as one of the manually coded facial expressions (classified as a 1) or not (classified as a 0). Employing multivariate binary logistic regression meant that a separate logistic regression model

was implemented for each of the manually coded facial expressions. The process for fitting a single model is outlined below:

1. Split the dataset into a train and test set. Here, all data for a single person must be contained in either the train or the test set (a father cannot be within both). This helps to avoid inflated prediction measures of generalization performance. We used 13 fathers in total: 10 of these comprised the training set ($n = 44{,}352$ frames), and 3 fathers comprised the test set ($n = 5{,}180$ frames).
2. Define our features $\boldsymbol{X}$ (the normalised FaceReader expression intensities) and our target variable $\boldsymbol{Y}$ (the manually coded facial expression). $\boldsymbol{X}$ is an $8 \times n$ matrix, where 8 is the number of FaceReader expressions, and n is the number of entries in the training dataset. $\boldsymbol{Y}$ is a binary array (1 indicates the manual facial expression of interest, 0 indicates any other expression) of length n, where n is the number of entries in the training dataset.
3. Using the Python package sk-learn, fit the logistic regression model on the training data. We used the LBFGS solver (an optimisation algorithm approximating the Broyden-Fletcher-Goldfarb-Shanno algorithm, see [33]), and we weighted the classes based on their frequencies in the training data, adjusting for the imbalance of behaviours per class (see Fig. 1).
4. Test the fitted model using the test set.

The fathers in the test and train datasets were selected through a trial-and-error process, with the aim of retaining a similar percentage of data per expression in each dataset, subject to the requirement of having all data from a single participant in only one dataset. The resulting representation of each manually coded expression in the full dataset, the training set, and the testing set is shown in Fig. 1 below.

We produced similar representation across almost all retained expressions (except for Mock Surprise, which was more heavily weighted in the test set). Following this evaluation, we excluded prediction models for Disgust, Surprise and Woe face, as these expressions each accounted for $< 1\%$ of the data.

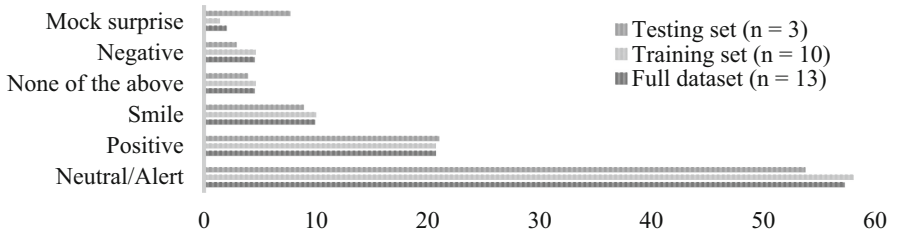


Fig. 1. Percentage of class occurrence in each dataset (%). Surprise, Woe face and Disgust were removed due to low prevalence ($< 1\%$).

3 Results

3.1 Quantifying FaceReader Performance Compared to Manual Coding

To address aim (1), we calculated how frequently the FaceReader software found the participant's face compared to the human coder. Our findings are provided in the contingency table below. Throughout our discussion, we assume the manually coded expressions to be correct, such that the manual expression is used as a benchmark with which the FaceReader output is compared.

Table 4. Face found vs. not found by the FaceReader software and the manual coder (%).

Manual	FaceReader (%)	
	Face Found	Face not found
Face found	10.71	31.84
Face not found	0.47	57.00

We found that throughout the 15,420 s of data, the manual coder identified a facial expression 42.55% of the time (from the table, manual face found = 10.71 + 31.84), while FaceReader only identified a facial expression 11.18% of the time (from the table, FaceReader face found = 10.71 + 0.47). Additionally, when a face was actually present (and was therefore coded by the human), FaceReader found the face 25.17% of the time (calculated by face found by both / total manual face found = 10.71 / (10.71 + 31.84) = 10.71 / 42.55). Table 4 also shows that percentage of time where FaceReader found a face but the Manual coder did not is very low (0.47%).

The observations where both FaceReader and the human coder classified a facial expression comprise the datasets used for the logistic regression analyses, i.e., we did not include data where the face was not found by either FaceReader or the human coder. The observations coded by both the human and the software comprise 1651 s of data, or 49,532 frames approximately.

3.2 The Relationship Between the Automated and Human Coding of Facial Expressions

To address aim (2), we looked to understand how accurately FaceReader facial expression intensities predicted manually coded facial expressions. The methodology for these analyses has been outlined in Sect. 2.4.

We fit six binary logistic regression models (one for each expression) to the training set, before testing the models on the test set. The resulting accuracy, sensitivity and specificity measures are provided in Table 5.

Accuracy represents the proportion of correct predictions for both classes; high accuracy means that the number of correct predictions (of either a 1 – the expression is present – or a 0 – the expression is not present) is high. Our six models all showed accuracy greater than 0.67, with most being greater than 0.80. The most accurate models were

Table 5. Accuracy, Sensitivity and Specificity for the logistic regression models.

Class	Predictive Performance Measure		
	Accuracy	Sensitivity	Specificity
Neutral ($n = 25{,}713$)	0.78	0.97	0.51
Positive ($n = 9172$)	0.67	0.48	0.72
Negative ($n = 2099$)	0.82	0.83	0.82
Smile ($n = 4437$)	0.94	0.84	0.94
None of the above ($n = 2021$)	0.88	0.27	0.91
Mock surprise ($n = 606$)	0.94	0.47	0.97

Mock Surprise and Smile (0.94), followed by None of the Above (0.88) and Negative (0.82). The prediction model for Positive (0.67) was the least accurate, followed by Neutral/Alert (0.78).

Sensitivity represents the model's ability to predict a true positive (i.e., correctly classifies a facial expression as present); high sensitivity means that the model rarely incorrectly classifies an expression as not present, and performs well at correctly classifying a facial expression as present). The sensitivity for some models were very high, including for Neutral (0.97), Smile (0.84) and Negative (0.83). However, the other models performed poorly: sensitivity for None of the Above was very low (0.27), with Mock Surprise (0.47) and Positive (0.48) only performing slightly higher.

Specificity represents the model's ability to predict a true negative (i.e., correctly class a facial expression as not present); high specificity means that the model rarely incorrectly classifies an expression as present), and performs well at correctly classifying facial expressions as not present). Our specificities were nearly all greater than 0.80. The highest specificity was for Mock Surprise (0.97), followed by Smile (0.94), None of the Above (0.91), Positive (0.82) and Negative (0.82). The lowest specificity was for Neutral/Alert (0.51).

4 Discussion

4.1 Summary of Results

Our work used 36 videos of fathers taken using headcams worn by infants during parent-infant interactions, totaling 15,420 s of data (after pre-processing). We manually coded the videos according to an established facial expression coding scheme [29], and also using an automated facial coding software. We applied contingency analysis to calculate how frequently FaceReader detected a face at the same time the human coder did, and used multivariate logistic regression models to evaluate the relationship between automated facial classification and manual coding.

Our results showed that FaceReader only found a facial expression around a quarter of the time that a human coder did (25.17%). This is not surprising, as it has been established that automated facial recognition is disadvantaged in real-world conditions

[13]. As such, we regard this low FaceReader performance as indicative of the high ecological validity of the interactions. Whilst previous studies have excluded data that FaceReader had failed to analyse, for reasons of signal loss during facial recognition [34], it was important that we retained these data in order to validate the software for naturalistic observations.

Our logistic regression models found that FaceReader predicted the manually coded expressions with generally good accuracy (*mean* = 0.84, *range* = 0.67–0.94), sensitivity (*mean* = 0.64, *range* = 0.27–0.97), and specificity (*mean* = 0.81, *range* = 0.51–0.97). The high accuracies across the six models suggest that our models were good at predicting when expressions were present. FaceReader was most accurate for Mock Surprise (0.94) and Smile (0.94). This is similar to [11], who found that FaceReader most accurately detected Surprise compared to other expressions. We found that FaceReader was least accurate for Positive (0.67) and Neutral/Alert (0.78), a finding inconsistent with other studies [5, 11, 13]. This is particularly interesting in the case of [13], who trialed FaceReader for in-the-wild facial detection, similarly to us.

We found high sensitivities for the Neutral/Alert, Smile and Negative models, meaning that FaceReader was able to positively identify the presence of these expressions. Conversely, FaceReader was less accurate in identifying None of the Above or Positive. This could be caused by a high number of false negative predictions, i.e., failing to predict an expression that was in fact present. This makes sense in the case of None of the Above, which serves to represent an "other" category, meaning that it encompassed multiple different expressions. In our work, this was mostly fathers eating, sneezing, yawning, or opening their mouths to mimic eating food, but could have included an even wider range of expressions. As such, we would not expect sensitivity to be as high for None of the Above as for the other expressions. For Mock Surprise and Positive, however, low sensitivity indicates that for some reason, our models often failed to correctly identify these expressions.

We generally found very high specificities (all were greater than 0.80 except for Neutral/Alert), meaning that FaceReader performed well at correctly identifying that a particular expression was absent. The low specificity for Neutral/Alert (0.51) indicates that the false positive rate was high – i.e., FaceReader incorrectly identified many expressions as Neutral. This is similar to [35], who reported that their low accuracy for Neutral (19%) was due to FaceReader over-detecting neutral expressions.

4.2 Failures in Face Detection

Previous work has highlighted the importance of creating more naturalistic settings for FaceReader applications [36]. In practice, naturalistic settings are problematic for a successful FaceReader analysis. In our work, the software struggled to detect a face across many scenarios where the human coder had no trouble. We looked to our videos to identify the reasons for this, and provide some examples here. Figure 2 shows some specific reasons that we believe FaceReader performance was low in our videos, including: blurry images (a), bad lighting (b), the face being partially out of shot (c and h), headcams/toys/food blocking the face (d and f), the parent facing away from the camera (e), or FaceReader misclassifying another object as a face (g). For confidentiality reasons, we are not able to share images of the fathers used within this study. However, the images

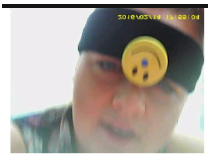
in Fig. 2 show a mother from a mother-specific, but otherwise identical, headcam study (this mother consented to have her data shared).

FaceReader documentation [31] outlines that the software performs less well if: the participant is wearing glasses, the lighting in the room is too dark, the participant is rotated away from the camera, or something is partially blocking the face (e.g., thick facial hair, hands, or a hat). Issues with glasses, bodily occlusions and artificial illumination have also been reported by others [13, 37]. The videos in our work were collected for an earlier study, so we were not able to advise that participants avoided wearing hats or glasses, that they sat in a room with natural lighting, or that they restrained from blocking their face with objects. It is therefore unsurprising that we observed a lot of these issues in the videos. Of our 13 fathers, six had facial hair of varying thickness, and three wore glasses. Additionally, many interactions also took place in front of a window, in rooms that were artificially lit, or in dim early morning or evening light, which lead to problems with lighting on fathers' faces (see Fig. 2b). It is possible that these factors lessened FaceReaders ability to analyse the faces in our videos.

While there are many reasons why FaceReader struggles to locate the face (Fig. 2), we suggest that the types of interactions we studied may also have affected this. In the free play interactions ($n = 10$), there were fast movements and variation between different positions (e.g., sit, lie on front), which led to images being both sporadic and blurry (see Fig. 2a). This also meant that the infant was not always at eye level with the father, meaning that the top, bottom, or side of the face was often out of shot (see Fig. 2c). Further, even if the full face was in sight, sometimes toys used for play would block the view of the face. While a human coder may be able to distinguish a facial expression in spite of small obstructions, slightly blurry shots, and missing parts of the face, FaceReader would not.

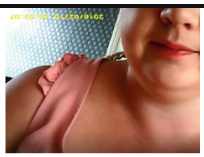
Similarly, the context of a Feeding interaction ($n = 24$) meant that cutlery, bowls, or food were often raised in front of the infant's face, partially or wholly blocking view of the father (see Fig. 2d). Further, while feeding interactions generally involved fathers sitting facing their infant in order to feed them, there were instances where the father was also eating a meal. This led to sideways or otherwise indirect views of the father (see Fig. 2e), if he was sitting next to the infant or somewhere around the table. While a human coder may distinguish a facial expression from a sideways or angled view, FaceReader is not able to do this beyond around a 40-degree tilt [31].

We also suggest that the use of the headcam led to a reduced FaceReader performance. For example, one problem we encountered was that the headcam was placed too high on the infant's head (i.e., pointing more upwards than forwards), so that only the top of the father's head was visible. Conversely, a headcam that was placed too low on the father's face meant that his eyebrows were covered, causing the face to be uncaptured by FaceReader (see Fig. 2f). In both of these cases, the face was typically still visible to the human coder. In a previous study investigating headcam use for capturing dyadic interactions [38], the experimenters manually adjusted the subject headcams to ensure that they were well fitted, and positioned to capture the most desirable perspective. In our work, subjects put the headcams on themselves, which meant that there was more likely to be placement issues.



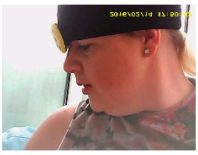
a. The image is blurry. This may happen if the infant (or father) moves their head too quickly.

b. The lighting in the room is bad, e.g., the parent is sitting in front of a window, or there is a lack of natural light.



c. Any part of the face is missing from the image. This may happen if the parent moves from the infant's line of sight, or the infant looks elsewhere.

d. A hand, toy or food-related item is blocking the view of the parent's face.



e. The parent is facing sideways, upwards, or downwards. This can obstruct facial features from view.

f. The headcam is blocking the view of the face, i.e., by being worn too low on the head, and covering the eyebrows.

g. A face that is not the parent's face is recognised and classified, (e.g., a face on a t-shirt, a stuffed animal, a painting, or another person.)

h. The parent is too close. This may happen when the parent leans in (e.g., to feed or kiss the infant). The face may also be unclassified if the parent is too far away.

Fig. 2. Examples of images where FaceReader does not find the face. Images are taken from an infant headcam during mother-infant interactions in a near identical study.

Finally, FaceReader also provided outputs for some frames that the human coder did not (0.47%). There are many reasons that this could happen, including where FaceReader misclassifies: a different person's face, an item of clothing or poster with a face on, or some other background object (see Fig. 2g). Complex backgrounds have previously been reported as problematic [13]. To account for this effect, we had already removed data from our analyses that included another caregiver in the background of the video, meaning that these misclassifications were likely to be caused by clothing or background items.

4.3 Implications for Future Work

To our knowledge, no previous studies have used FaceReader to analyse headcam videos. It would therefore be useful to further evaluate how headcams can be best used to optimise the chances of a successful FaceReader analysis. This is particularly important as the use of headcams allows for a more ecologically valid and natural observation [14]. We therefore outline six recommendations for future work in this area, aiming to maintain

the authenticity of a natural observation, while also optimising the logistical aspects of the observation.

[1] Subjects should be instructed on how to properly put the headcams on both themselves and the infant (to avoid pointing up/down and missing the partners face).
[2] Subjects should be advised regarding optimal lighting conditions (i.e., natural light). Where possible, observations should take place in an area of natural lighting, preferably during the middle of the day where natural light is most prevalent.
[3] Subjects should be advised regarding glasses (not to wear them if possible), or other facial occlusions. Whilst no hands in front of the face would be desirable, highlighting this could affect the authenticity of subject behaviours and movements, so we would not recommend mentioning this to subjects.
[4] Depending on the interaction, researchers could advise subject posture (e.g., we found that feeding interactions had successful analysis when subject was face to face with infant). It would be beneficial to suggest interactions or observations that naturally cause subjects to face the camera front-on.
[5] Subjects should not sit in view of photos or posters hung on walls, and should not wear t-shirts or other clothing items with people shown on them.
[6] Researchers should support the development and usage of more powerful compact cameras, to provide greater robustness against rapid head movements (as previ-ously suggested by [9])

While these recommendations should aid in optimising FaceReader performance for naturalistic interactions, the reality is that – for now – it may be necessary that manual coding (or some other method) is used to supplement FaceReader coding. In our case, this would account for roughly 25% of faces being automatically coded, and the remaining having to be supplemented by a human coder (this is, if the end goal is 100% coding). However, it may be that this is sufficient for drawing useful conclusions in many cases, especially when using high quality, long extracts of video data. While this is not nearly close to the goal of fully automated facial coding, 25% represents a starting benchmark which future work can aim to improve upon.

Additionally, FaceReader provides much more detail than what is capable from a human coder, i.e., expression intensity for eight concurrent expressions. This means that even with a performance rate of 25%, we can potentially learn a lot more from the FaceReader output that wouldn't be possible from manual coding alone. In a clinical scenario, therefore, where manual coding is impractical, it is easy to see the potential utility of automated coding techniques, even at a 25% performance rate.

Future work could also identify whether the successful automated coding is biased towards certain expressions (e.g., expressions may be more prevalent when FaceReader is unsuccessful, such as when the second caregiver is present, or when the parent or infant turns toward a distraction). Similarly, it would be beneficial to investigate how FaceReader could work alongside complementary techniques (e.g., linear interpolation) to identify "missing" facial expressions.

4.4 Strengths and Limitations

A major strength of our research is the use of real-life observations, which meant that fathers exhibited natural, unposed facial expressions. Without the presence of a third-party researcher to record the interactions, it is likely that the recordings captured more ecologically valid facial expressions [14]. Also, using headcams during a dyadic interaction allowed the camera to focus directly on fathers' faces. This is an advantage over third person cameras, which can miss out on capturing facial expressions [14].

For limitations, we acknowledge there was low prevalence of some facial expressions, meaning that some had to be excluded from the logistic regression models (e.g., Woe face, Disgust). Additionally, some of the models possibly did not perform as well as they might have with more data. The facial expressions are a direct reflection of fathers' emotions while interaction with their children, however, so it is not surprising that we did not encounter lots of Disgust, for example.

Further, there was not a one-to-one relationship between the facial expressions in the manual coding scheme and those implemented by FaceReader. For example, the manual coding scheme contained the expression Negative, while FaceReader contained the separate negative expressions Sad, Scared, Angry and Contempt. If we had manually-coded these negative expressions separately, it is possible that we may have been able to implement better performing, distinct models for each expression. Having said that, the prevalence of Negative was quite low ($n = 2099$) compared to other expressions (e.g., Neutral/Alert, Smile), suggesting that there may not have been quite enough data to have created three, high performing, separate models in this instance.

As previously acknowledged, the length of video material for each subject was not the same (i.e., we had many more facial expressions for some fathers than others, especially where some fathers provided multiple videos). It is possible that this led to biases in how our models learned to predict certain facial expressions. However, we modulated for this effect in our performance measures by including all data from a given subject in either the training or the testing dataset.

There are some inherent biases present within the ALSPAC cohort, many of which have been detailed previously [26]. One example is that the cohort is mostly of White-European origin, reducing the generalisability of findings to the general population. However, as our work does not aim to interpret specific facial expressions present in the interactions, generalisability of findings is not as important here as for other studies.

Finally, we acknowledge that wearing headcams might have influenced the natural facial expressions of our participants. Although, we believe that the effect of this was mitigated (and countered) by the more ecologically valid expressions that we expected to see from real life, at-home interactions.

Acknowledgements. We are extremely grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses. The UK Medical Research Council and Wellcome (Grant ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. We would also like to thank the two double coders who were involved in this work, Lottie Relph and Maddy Stephens.

This publication is the work of the authors RB, IC, HB, and RP, who will serve as guarantors for the contents of this paper. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf); This work is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 758813; MHINT).

Additionally, RB was supported by the Engineering and Physical Sciences Research Council (EPSRC) Digital Health and Care Centre for Doctoral Training (CDT) at the University of Bristol (UKRI Grant No. EP/S023704/1). IC was supported by the Wellcome Trust Research Fellowship in Humanities and Social Science (Grant ref: 212664/Z/18/Z).

Appendix

Here we provide a flow diagram to demonstrate how we processed the data involved in this study.

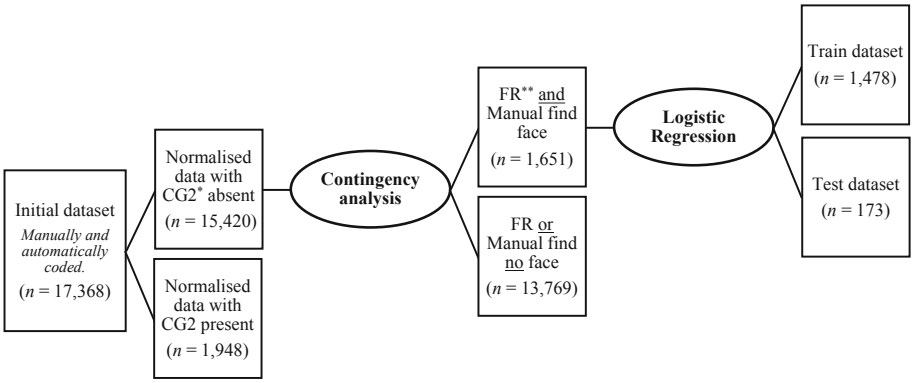


Fig. 3. Flow diagram to demonstrate stages of data processing; *n* refers to the number of video frames. *CG2 = Caregiver 2. **FR = FaceReader.

References

1. Noldus: FaceReader (2022). https://www.noldus.com/facereader
2. Den Uyl, M.J., Van Kuilenburg, H.: The FaceReader: online facial expression recognition. In Proceedings of Measuring Behavior, vol. 30, no. 2, pp. 589–590. Wageningen (2005)
3. Lewinski, P., den Uyl, T.M., Butler, C.: Automated facial coding: validation of basic emotions and FACS AUs in FaceReader. J. Neurosci. Psychol. Econ. **7**(4), 227 (2014)
4. Skiendziel, T., Rösch, A.G., Schultheiss, O.C.: Assessing the convergent validity between the automated emotion recognition software Noldus FaceReader 7 and facial action coding system scoring. PLoS ONE **14**(10), e0223905 (2019)
5. Terzis, V., Moridis, C.N., Economides, A.A.: Measuring instant emotions based on facial expressions during computer-based assessment. Pers. Ubiquit. Comput. **17**(1), 43–52 (2013)

6. Terzis, V., Moridis, C.N., Economides, A.A.: Measuring instant emotions during a self-assessment test: the use of FaceReader. In: Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research, pp. 1–4 (2010)
7. Talen, L., den Uyl, T.E.: Complex website tasks increase the expression anger measured with FaceReader online. Int. J. Human–Comput. Interact. 1–7 (2021)
8. Zaman, B., Shrimpton-Smith, T.: The FaceReader: measuring instant fun of use. In: Proceedings of the 4th Nordic conference on Human-Computer Interaction: Changing Roles, pp. 457–460 (2006)
9. Danner, L., Sidorkina, L., Joechl, M., Duerrschmid, K.: Make a face! Implicit and explicit measurement of facial expressions elicited by orange juices using face reading technology. Food Qual. Prefer. **32**, 167–172 (2014)
10. Benţa, K.I., et al.: Evaluation of a system for realtime valence assessment of spontaneous facial expressions. In: Distributed Environments Adaptability, Semantics and Security Issues International Romanian-French Workshop, Cluj-Napoca, Romania , pp. 17–18 (2009)
11. Brodny, G., Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., Wróbel, M.R.: Comparison of selected off-the-shelf solutions for emotion recognition based on facial expressions. In: 2016 9th International Conference on Human System Interactions (HSI), pp. 397–404. IEEE (2016)
12. Krishna, T., Rai, A., Bansal, S., Khandelwal, S., Gupta, S., Goyal, D.: Emotion recognition using facial and audio features. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 557–564 (2013)
13. Gómez Jáuregui, D.A., Martin, J.C.: Evaluation of vision-based real-time measures for emotions discrimination under uncontrolled conditions. In: Proceedings of the 2013 on Emotion Recognition in the Wild Challenge and Workshop, pp. 17–22 (2013)
14. Lee, R., et al.: Through babies' eyes: practical and theoretical considerations of using wearable technology to measure parent–infant behaviour from the mothers' and infants' viewpoints. Infant. Behav. Dev. **47**, 62–71 (2017). https://doi.org/10.1016/j.infbeh.2017.02.006
15. Karreman, A., Riem, M.M.: Exposure to infant images enhances attention control in mothers. Cogn. Emot. **34**(5), 986–993 (2020)
16. Lyakso, E., Frolova, O., Matveev, Y.: Facial Expression: psychophysiologcal study. In: Handbook of Research on Deep Learning-Based Image Analysis Under Constrained and Unconstrained Environments, pp. 266–289. IGI Global (2021)
17. O'Brien, M.: Shared caring: bringing fathers into the frame (2005)
18. Tamis-LeMonda, C.S., Shannon, J.D., Cabrera, N.J., Lamb, M.E.: Fathers and mothers at play with their 2-and 3-year-olds: Contributions to language and cognitive development. Child Dev. **75**(6), 1806–1820 (2004)
19. Ramchandani, P.G., Domoney, J., Sethna, V., Psychogiou, L., Vlachos, H., Murray, L.: Do early father–infant interactions predict the onset of externalising behaviours in young children? Findings from a longitudinal cohort study. J. Child Psychol. Psychiatry **54**(1), 56–64 (2013)
20. Feldman, R.: Infant–mother and infant–father synchrony: the coregulation of positive arousal. Infant Mental Health J. Official Publ. World Assoc. Infant Mental Health **24**(1), 1–23 (2003)
21. Montague, D.P., Walker-Andrews, A.S.: Peekaboo: a new look at infants' perception of emotion expressions. Dev. Psychol. **37**(6), 826 (2001)
22. Kokkinaki, T., Vasdekis, V.G.S.: Comparing emotional coordination in early spontaneous mother–infant and father–infant interactions. Eur. J. Develop. Psychol. **12**(1), 69–84 (2015)
23. Harris, P.A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., Conde, J.G.: Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J. Biomed. Inform. **42**(2), 377–381 (2009). https://doi.org/10.1016/j.jbi.2008.08.010

24. Boyd, A., et al.: Cohort profile: the 'children of the 90s'; the index offspring of the avon longitudinal study of parents and children (ALSPAC). Int. J. Epidemiol. **42**, 111–127 (2013). https://doi.org/10.1093/ije/dys064
25. Fraser, A., et al.: Cohort profile: the avon longitudinal study of parents and children: ALSPAC mothers cohort. Int. J. Epidemiol. **42**, 97–110 (2013). https://doi.org/10.1093/ije/dys066
26. Lawlor, D.A., et al.: The second generation of the Avon longitudinal study of parents and children (ALSPAC-G2): a cohort profile. Wellcome open research, 4, 36 (2019). https://doi.org/10.12688/wellcomeopenres.15087.2
27. Northstone, K, et al.: The Avon longitudinal study of parents and children (ALSPAC): an update on the enrolled sample of index children in 2019. Wellcome Open research, 4:51 (2019). https://doi.org/10.12688/wellcomeopenres.15132.1
28. Noldus. The Observer XT (2022a). http://www.noldus.com/human-behavior-research/products/the-observer-xt
29. Costantini, I., et al.: Mental health intergenerational transmission (MHINT) process manual (2021). https://doi.org/10.31219/osf.io/s6n4h
30. Gudi, A.; Tasli, H.E.; Den Uyl, T.M.; Maroulis, A.: Deep learning based facs action unit occurrence and intensity estimation. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) 4 May 2015, vol. 6, pp. 1–5. (2015)
31. Loijens, L., Krips, O., Grieco, F., van Kuilenburg, H., den Uyl, M., Ivan, P.: FaceReader 8 reference manual, noldus information technology (2020)
32. Van Rossum, G., Drake, F.L.: Python 3 reference manual. scotts valley, CA: CreateSpace (2009)
33. Fletcher, R.: Practical Methods of Optimization. John Wiley & Sons, Hoboken (2013)
34. Weth, K., Raab, M.H., Carbon, C.C.: Investigating emotional responses to self-selected sad music via self-report and automated facial analysis. Music. Sci. **19**(4), 412–432 (2015)
35. Matlovic, T., Gaspar, P., Moro, R., Simko, J., Bielikova, M.: Emotions detection using facial expressions recognition and EEG. In: 2016 11th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), pp. 18–23. IEEE (2016)
36. Booijink, L.I.: Recognition of emotion in facial expressions: the comparison of FaceReader to fEMG and self-report (Master's thesis) (2017)
37. Webber, M.: Can jealousy be detected as a unique pattern of recordable facial expressions by the FaceReader, and thus do such expressions manifest differently between sexes upon exposure to jealousy–evoking Snapchat messages?" (2018)
38. Park, C.Y., et al.: K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. Sci. Data **7**(1), 1–16 (2020)