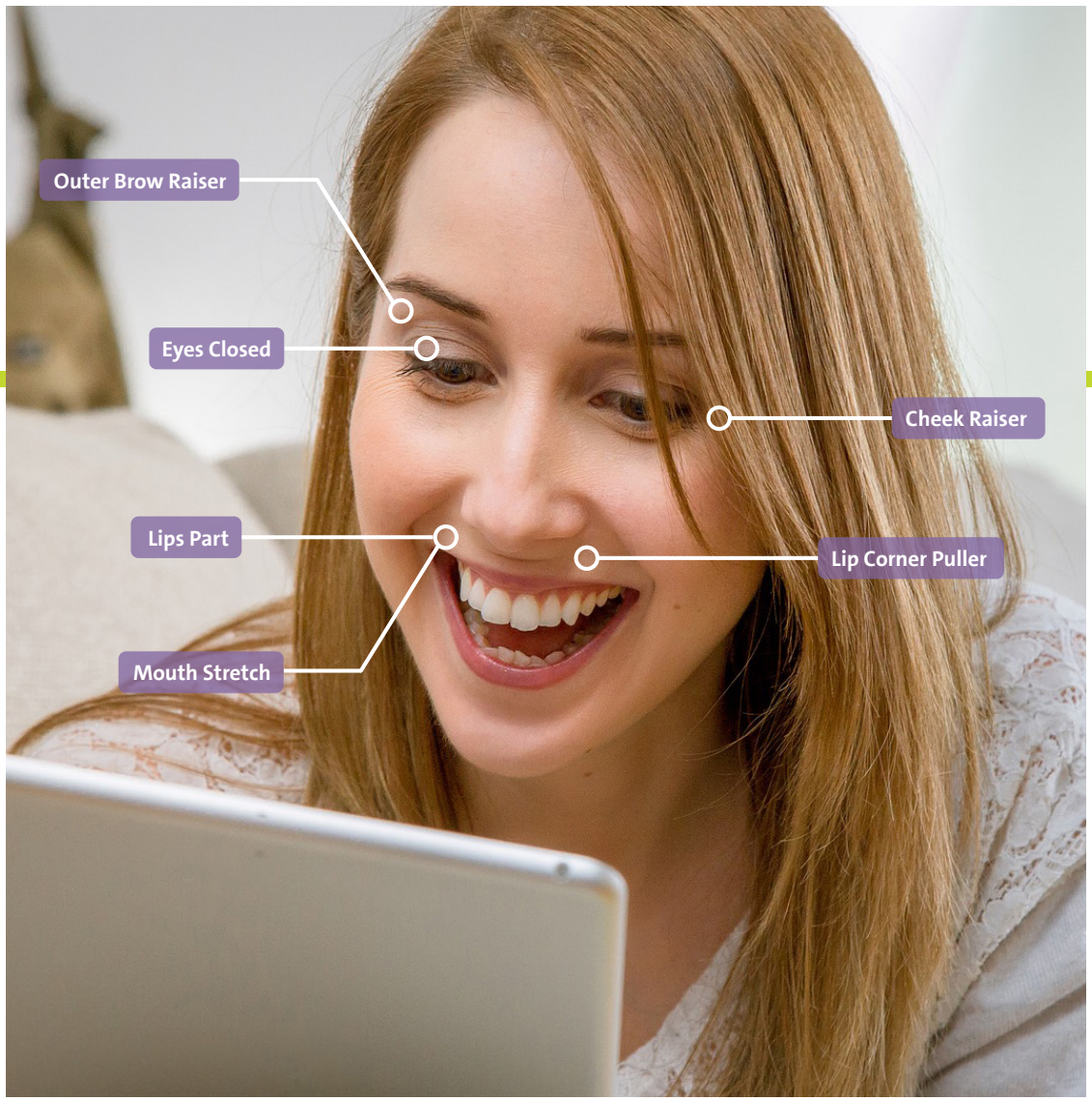


Validation Action Unit Module

By Amogh Gudi and Paul Ivan, Vicarious Perception Technologies

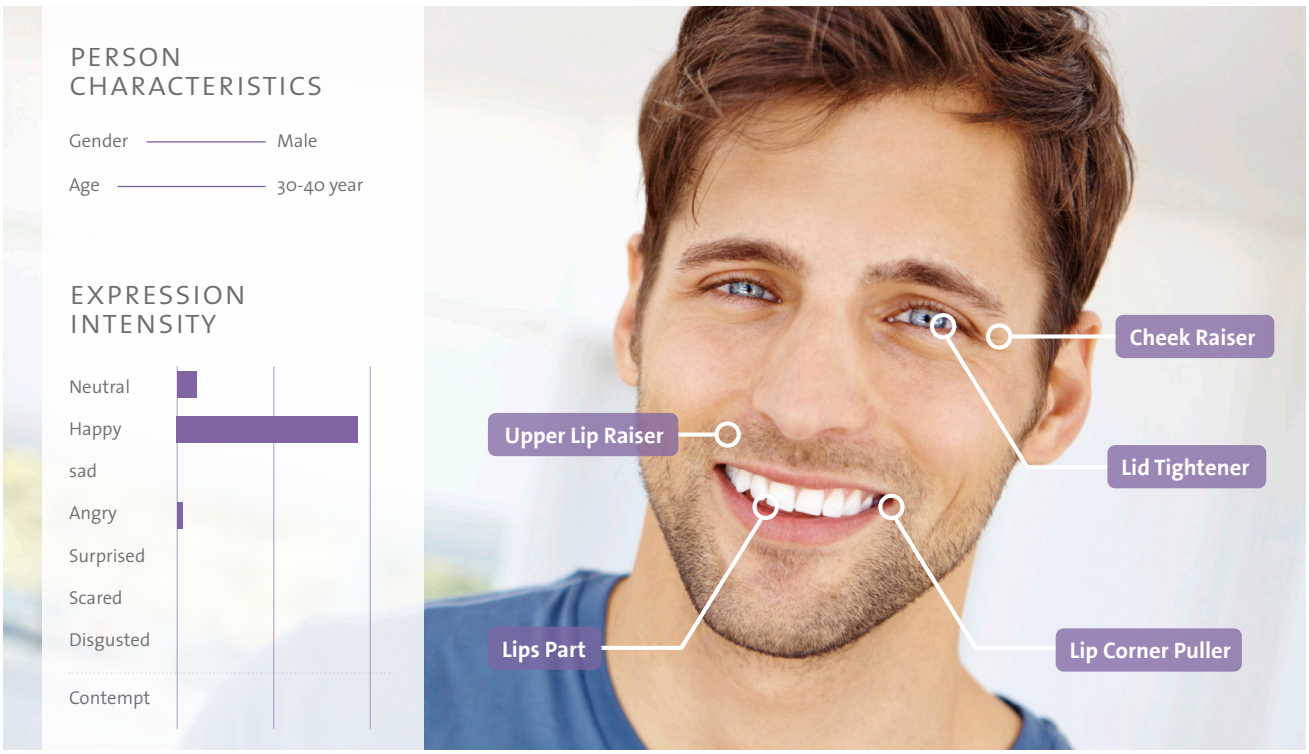
A white paper by Noldus Information Technology

INTRODUCTION

This document presents a validation of the FaceReader 9 Action Unit Module by evaluating the performance on an annotated dataset. It provides the agreement between the FaceReader classifications and the coded data-set, as well as evaluation of the performance of the individual AU classifiers.

DATASET

The dataset used for this validation is the Amsterdam Dynamic Facial Expression Set (ADFES) [1] that consists of 22 models (10 female, 12 male) performing nine different emotional expressions (anger, disgust, fear, joy, sadness, surprise, contempt, pride, and embarrassment). As there was no FACS [2] coding available for this dataset, we had a selection of the data-base (213 images) annotated manually by two certified FACS coders.



AGREEMENT

To assess the reliability of the annotated dataset, we can calculate the agreement between the two FACS coders using the Agreement Index, as described by Paul Ekman *et al.* in the FACS Manual [2].

FORMULA

This index can be computed for every annotated image according to the following formula:

*(Number of AUs that both coders agree upon) * 2 / The total number of AUs scored by the two coders*

The agreement between FaceReader AU classifications and the human-coded dataset was 0.81.

EXAMPLE

If an image was coded as 1+2+5+6+12 by one coder and as 5+6+12 by the other, the agreement index would be: 3 * 2 / 8 = 0.75. Note that the intensity of the action unit (AU) classification is ignored for the calculation of the agreement index, the focus is on the AU being active or not.

AGREEMENT

The agreement between the two human FACS coders on the selection of images from the ADFES dataset was 0.83. A minimum agreement score of 0.80 is usually regarded as reliable [2], so this minimum is met for the annotation of this database. The agreement between the FaceReader AU classifications and the human-coded dataset was 0.81. Agreement index is the non-normalized F1 score over the whole dataset.

To pass the official FACS certification test [3,4,5], an agreement score of 0.70 or above should be obtained. This is achieved for a number of AUs, the next section describes the performance.

ACTION UNIT PERFORMANCE

Apart from the agreement between FaceReader (FR) and the database, it is also interesting to examine the accuracy of every individual action unit. Therefore, the table below shows the performance of 17 AUs that FR is capable of classifying and were also present in the dataset. The difficulty of determining the accuracy of AU classifiers, is that the obtained performance depends heavily on the chosen evaluation metric. Taking this into account, the table shows the AU performance on five common evaluation metrics, i.e. Recall, Precision, F1, and Accuracy.

Table. *ADFES (AU), 213 annotated images.*

AU	Present	Recall	Precision	F1	Accuracy
1	100	0,94	0,90	0,92	0,92
2	72	0,94	0,89	0,92	0,94
4	84	0,98	0,72	0,83	0,84
5	67	0,88	0,78	0,83	0,88
6	53	0,91	0,77	0,83	0,91
7	58	0,53	0,60	0,56	0,77
9	22	0,95	0,95	0,95	0,99
10	16	0,75	0,52	0,62	0,93
12	58	0,93	0,74	0,82	0,89
14	54	0,78	0,84	0,81	0,91
15	26	0,81	0,78	0,79	0,95
17	65	0,83	0,78	0,81	0,88
20	21	0,95	0,49	0,65	0,90
23	15	0,73	0,50	0,59	0,93
24	26	0,92	0,77	0,84	0,96
25	86	0,97	0,91	0,94	0,95
26	28	1,00	0,34	0,50	0,74
Average	**50,1**	**0,87**	**0,72**	**0,78**	**0,90**

EVALUATION METRICS

A brief description for the terms used in the table:

- *AU* – the action unit number.
- *Present* – the number of times an AU was coded in the dataset.
- *Recall* – denotes the ratio of annotated AUs that were detected by FaceReader. A recall of 0.94 for AU1 indicates, for example, that 94% of the annotated images with AU1 are classified as such by FaceReader [6].
- *Precision* – a ratio denoting how often FaceReader is correct when classifying an AU as present. For example, in the case of AU1 the FaceReader classification is correct 90% of the time [6].
- *F1* – there exists a trade-off between the recall and precision measures, and a good classifier ought to have a decent score on both measures. The F1 measure summarizes this trade-off in a single value and is computed using the formula: 2 * ((precision * recall) / (precision + recall)) [7].
- *Accuracy* – simply represents the percentage of correct classifications. It is computed by dividing the number of correctly classified images (both positive and negative) by the total number of images [8].

ACTION UNIT EVALUATION

> The F1 measure combines the important recall and precision measures, and displays the largest differences between AU classifiers.

To evaluate and compare the quality of the individual AU classifiers from the data in the table above, we could make a discrimination between the AU classifiers based on the F1 measure. The F1 measure is a suitable metric for this purpose because it combines the important recall and precision measures, and displays the largest differences between the AU classifiers. Based on the F1 measure the best classifiers – those that might be good enough already to pass the FACS test – are AUs 1, 2, 4, 5, 6, 9, 12, 14, 15, 17, 24, and 25 (F1: 0.79 - 0.95); the classifiers that performed reasonably well are AUs 7, 10, 20, and 23 (F1: 0.56 - 0.65); and the only AU that performed not so well is AU26 (F1: 0.50). FaceReader can also classify AUs 18, 27 and 43, but these were not sufficiently present in the dataset to evaluate their performance.

BABY FACEREADER

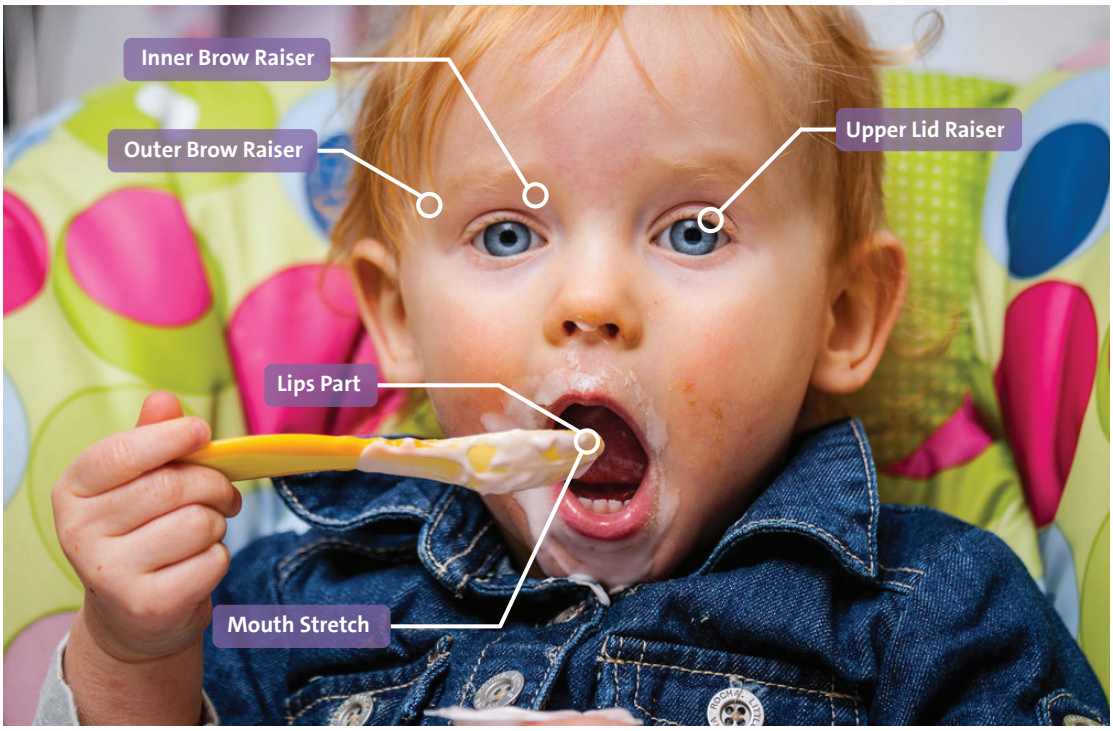
For validation of Baby FaceReader the Baby FACS Manual Dataset was used, which consists of 67 annotated images [9]. The table below shows the performance of the analysis of Action Units in Baby FaceReader.

Table. *Baby FACS Manual (AU), 67 annotated images. See Evaluation Metrics on Page 5 for a brief description for the terms used in the table.*

AU	Present	Recall	Precision	F1	Accuracy
1	13	0,92	0,40	0,56	0,77
2	14	0,43	0,43	0,43	0,80
3+4	12	1,00	0,48	0,65	0,84
5	10	0,80	0,50	0,62	0,88
6	39	0,97	0,64	0,78	0,73
7	24	0,75	0,60	0,67	0,78
12	13	1,00	0,48	0,65	0,83
15	11	0,73	0,67	0,70	0,91
17	17	0,76	0,54	0,63	0,81
20	13	0,92	0,46	0,62	0,81
25	54	0,91	0,94	0,92	0,90
26	25	0,84	0,60	0,70	0,78
27	16	0,75	0,71	0,73	0,89
43	18	1,00	0,60	0,75	0,85
Average	**19,9**	**0,84**	**0,58**	**0,67**	**0,83**

REFERENCES

1. Van der Schalk, J.; Hawk, S. T.; Fischer, A. H.; Doosje, B. J. (2011). Moving faces, looking places: The Amsterdam Dynamic Facial Expressions Set (ADFES). *Emotion*, **11**, 907-920. DOI: 10.1037/a0023853
2. Ekman, P.; Friesen, W. V.; Hager, J. C. (2002). *Facial action coding system: The manual on CD-ROM. Instructor's Guide*. Salt Lake City: Network Information Research Co.
3. https://www.erikarosenberg.com/is-facs-training-right-for-you
4. Lewinski, P., den Uyl, T. M., & Butler, C. (2014). Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader. *Journal of Neuroscience, Psychology, and Economics*, **7(4)**, 227.
5. Chapter 6: FACS Score Checker Computer Program from FACS Investigator's Guide. https://www.paulekman.com/facial-action-coding-system/
6. http://en.wikipedia.org/wiki/Precision_and_recall
7. http://en.wikipedia.org/wiki/F1_Score
8. http://en.wikipedia.org/wiki/Accuracy
9. Oster, H. (2016). *Baby FACS: Facial Action Coding System for infants and young children*. Unpublished monograph and coding manual. New York University.



INTERNATIONAL HEADQUARTERS
Noldus Information Technology bv
Wageningen, The Netherlands
Phone: +31-317-473300
Fax: +31-317-424496
E-mail: info@noldus.nl

NORTH AMERICAN HEADQUARTERS
Noldus Information Technology Inc.
Leesburg, VA, USA
Phone: +1-703-771-0440
Toll-free: 1-800-355-9541
Fax: +1-703-771-0441
E-mail: info@noldus.com

REPRESENTATION
We are also represented by a worldwide network of distributors and regional offices. Visit our website for contact information.

WWW.NOLDUS.COM