Systematic Review on Interrater Agreement in Facial Emotion Recognition Databases

AUTHOR: MANA MAHMOUDI
M.Mahmoudi-1@student.tudelft.nl
RESPONSIBLE PROFESSOR: BERND DUDZIK

TU Delft

1. BACKGROUND

Definitions :

- **Affect** : subjective feeling encompassing emotions, moods, etc.
- **Interrater agreement (IRA)** : consistency between two or more raters, annotators [1]
- **Automatic Facial Affect Recognition** : determining the affective state of a subject based on provided visual facial input
- **Affect Representation Schemes (ARS)** : a method for defining affective states

Motivation :

Recognizing facial emotions is key for social interaction, but emotion labeling is subjective and poses challenges for automatic facial affect prediction (see Fig. 1). While multiple raters and IRA measures are used, the extent of their use is not well-defined.

2. RESEARCH QUESTION

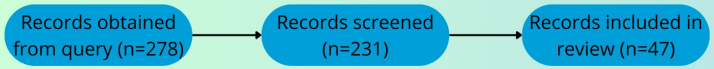
What are the differences in interrater agreement measurement methodologies among published datasets for automatic facial affect prediction?

Topics that will be covered:

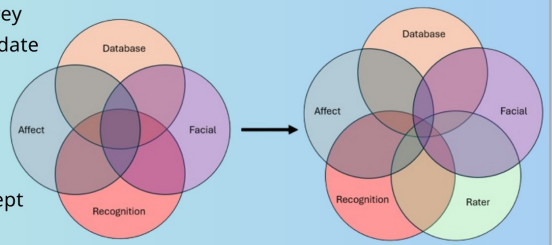
- Types of affective states
- Different affect representation schemes
- Interrater agreement methodologies

3. METHODOLOGY

- Systematic literature review ensures reproducibility of the survey
- PRISMA guidelines [2] are followed for structure for conducting and reporting systematic reviews
- Bibliograpic databases: Scopus, IEEE, ACM, Web of Science
- The concepts used for the queries are: *face, emotion, recognition, database* and *rater*.
- Obtained records were screened by Title and Abstract first then by Full Text.
- Data extraction was done while screening the full text for time efficiency reasons.



- Eligibility criteria:
  - Describes datasets with facial affect recognition annotations.
  - Introduces a previously unpublished dataset or provides new affective annotations to existing facial dataset.
  - Paper published in English
  - Paper is not a survey
  - Paper publication date ≤ April 2024
- Feasibility criteria:
  Adding the *rater* concept

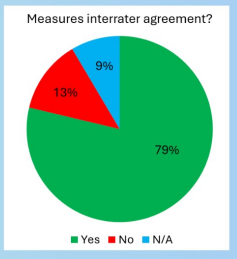


Narrowed down from 12 908 to 183 (on Scopus)

4. RESULTS

IRA measurement :

47 papers were retrieved, and a large majority of them measured IRA (79%).

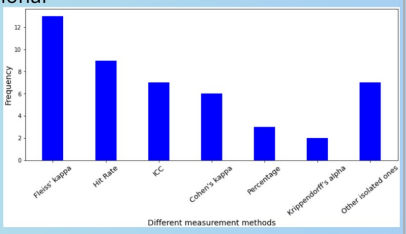


ARS and affect states :

22/47 papers had Ekman's basic emotions included in their ARS (see Fig. 2). 3 papers used dimensional type ARS only, and 6 used both dimensional and categorical.

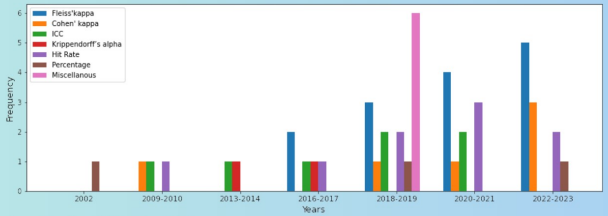
IRA methodologies :

Fleiss' kappa is the most used methodology, followed by Hit Rate.



IRA over time:

The data retrieved was from 2002 onwards. A trend of Fleiss' kappa growing can be noticed. Cohen's kappa has a peak on the last year range. Hit Rate and ICC keep quite a consistent presence over the years.



Strategies to facilitate IRA :

Improve faces' input or the selection of the annotators.

5. DISCUSSIONS & CONCLUSION

- No correlation was seen between ARS and IRA.
- Time provided it would be possible to retrieve more papers and stronger assumptions could be made.
- The use of IRA is prominent. Fleiss' kappa being the most used one suggests that the use of complex IRA methodologies is approved. They account for chance agreement and provide a more nuanced view of interrater reliability [3].
- The strategy to achieve high IRA by using "perfect" facial inputs may negatively impact the performance of systems trained on such a database.
- No low average IRA was found but the methodologies and techniques to compute IRA should be analyzed if disclosed by the paper.

6. FUTURE WORK

Analyze the relationship between the IRA of a published database and the empirical performance of machine learning systems trained on the database.
Does the level of IRA influence the real-world applications reliability?

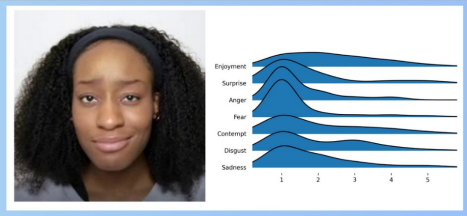


Fig 1. Ambiguity and uncertainty when labeling emotions [4]

Fig 2. Ekman's basic emotions - anger, fear, disgust, surprise, happiness, sadness [5]

REFERENCES

[1] Popping Roel. Introduction to Interrater Agreement for Nominal Data.
[2] Page M J, McKenzie J E, Bossuyt P M, Boutron I, Hoffmann T C, Mulrow C D et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews BMJ 2021; 372 :n71 doi:10.1136/bmj.n71
[3] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: The kappa statistic," Family medicine, vol. 37, no. 5, pp. 360–363, 2005.
[4] F. Cabitza, A. Campagner, and M. Mattioli, "The unbearable (technical) unreliability of automated facial emotion recognition," Big Data & Society, vol. 9, p. 20539517221129549, July 2022. Publisher: SAGE Publications Ltd.
[5] P. Ekman, "Basic emotions," in Handbook of Cognition and Emotion. John Wiley Sons, Ltd, 1999, ch. 3, pp. 45–60.