Systematic Review on Interrater Agreement in Facial Emotion Recognition Databases

Mana Mahmoudi[1]
Supervisor: Bernd Dudzik[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Mana Mahmoudi
Final project course: CSE3000 Research Project
Thesis committee: Bernd Dudzik, Catharine Oertel

Abstract

Recognizing facial emotions is key for social interaction, yet the subjective nature of emotion labeling poses challenges for automatic facial affect prediction. Variability in how individuals interpret emotions leads to uncertainty in training data for machine learning models. While multiple raters and interrater agreement (IRA) measures are used to address this, the extent of their use and their impact on dataset reliability is not well understood. This systematic literature review investigates the methodologies used to measure IRA in facial affect recognition datasets. Concrete eligibility and feasibility criteria were applied, and it resulted in 47 papers being retrieved from Scopus, Web of Science, IEEExplore, and ACM Digital Library. Data on affect states, affect representation schemes (ARS), and IRA methodologies used by the datasets and their corresponding papers were extracted to provide a comprehensive overview and allow a detailed analysis. Clear correlation was not found in between ARS and IRA, but the retrieved data showed that Fleiss' kappa was the most popular methodology over time but also in the recent years.

1 Introduction

There are technological advancements that rely on being able to adapt to and predict a user's emotion, particularly through facial expression recognition. Facial emotion recognition can be used, for instance, to increase safety in self-driving cars [1] or help in mental stress detection among university students [2]. Recognizing and distinguishing facial emotions is fundamental for effective interaction and social connection, as human facial expressions play a key role in communication across various social settings [3]. Despite the significant potential of automatic facial affect prediction, challenges remain, particularly in the area of emotion labeling.

This paper is divided into 6 sections. Following this introduction, section 2 outlines the methodology for this literature review. Section 3 presents the results found from the relevant set of papers. Section 4 discusses the reproducibility matters as well as the ethical considerations of the survey. In section 5, there will be a discussion of the results, while the conclusion and future work can be found in section 6. First, this introduction section is broken down in to a background at 1.1, then 1.2 for a review of related work, and lastly 1.3 for a detailed presentation of the research question and sub-questions.

1.1 Background

One of the primary difficulties in automatic facial affect prediction lies in the subjective nature of emotion labeling. Different individuals may interpret the same cues differently, causing uncertainty in the ground truth labels used for training machine learning models [4]. This subjectivity can significantly impact the reliability of emotion recognition systems and poses a barrier to their widespread adoption [5].

To address this challenge, research has explored the use of multiple raters and interrater agreement (IRA) measures to monitor the uncertainty and reliability of emotion labels [6]. IRA measures assess the extent to which different raters provide consistent labels for the same set of data, offering a way to quantify and potentially improve labeling reliability as it gauges data accuracy and representation [7]. However, there remains a gap in understanding the extent to which these measures are used across different automatic facial affect prediction datasets. Furthermore, the overall level of agreement among raters for these datasets and its impact on the performance of the technological system remains unclear.

1.2 Related Work

Previous research has investigated various aspects of automatic facial affect prediction, including the subjectivity of emotion labeling and the reliability of emotion recognition systems. For example, the paper from Cabitza et al. [4] highlights the technical unreliability of automated facial emotion recognition, emphasizing the need for more robust labeling techniques. Other studies have focused on the application of IRA measures to improve the consistency of emotion labels [8]. However, there is limited research systematically reviewing how these measures are implemented across different datasets.

This survey complements existing work by providing a comprehensive analysis of the methodologies used to measure IRA in published datasets. By identifying patterns and best practices, this paper aims to provide a foundation for future research in this area.

1.3 Research Question

By systematically reviewing existing literature, we can uncover underlying patterns, differences, and best practices that might not be immediately evident through a brief examination [9]. This methodological choice also enhances the transparency and reproducibility of the findings, ensuring they are robust and can be further explored. This survey paper focuses on filling this knowledge gap by systematically reviewing existing datasets and investigating the prevalence of IRA measures. The research question can be defined as such: **What are the differences in interrater agreement measurement methodologies among published datasets for automatic facial affect prediction?**. To answer this question, different sub-questions are made, as steps to objectively analyze the different datasets. The break down of the research question leads to five attainable objectives, presented as sub-questions:

- SQ1: What types of affective states have been targeted by datasets?
- SQ2: What different affect representation schemes have been used in these datasets?
- SQ3: Do datasets collect multiple ratings for a record (and how many)?
  - SQ3a: If so, do datasets measure interrater agreement?
  - SQ3b: What measures do they use for this (and what is the level of agreement)?
  - SQ3c: Do dataset creators use any strategies to facilitate interrater agreement (and what are these strategies)?
- SQ4: Is there a change in how datasets measure interrater agreement over time?
- SQ5: Is there a relationship between the affect representation scheme used by datasets and their interrater agreement?

With this knowledge gap filled, it will be possible to go further in the analysis of the relationship between the IRA in datasets and the affect prediction systems' empirical performance. This will further be discussed under the Conclusion and Future Work section.

2 Methodology

To perform a systematic literature review (SLR), the PRISMA guidelines (which stands for Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [10] are followed, allowing this review to be reproducible. An SLR is a structured approach to reviewing literature that involves systematically searching, filtering, and synthesizing all relevant studies on a particular topic, in contrast to other forms of literature review that may not be as comprehensive or systematic. According to Grant and Booth [11], various review types exist, but an SLR is particularly rigorous in its methodology.

This section outlines the stages of the SLR process: **Searching**: section 2.2 justifies the different search engines used. **Filtering**: section 2.1 presents the eligibility criteria for selecting papers. **Extraction**: section 2.3 details the strategy to obtain the required papers, and section 2.4 introduces the search process. The search results are presented in section 2.5.

2.1 Eligibility Criteria

Defining a scope is essential for selecting relevant papers. Inclusion and exclusion criteria help ensure that the review aligns with the research question and time constraints, following the PRISMA guidelines [10] as per the fifth item on the checklist: *Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses* [12]. The criteria were chosen to respect the research question topic and the time frame allocated for this review.

Table 1 provides a structured summary of these criteria, ensuring clarity and reproducibility in the selection process.

2.2 Search Engines

The bibliographic databases used to gather papers were: Scopus, Web Of Science, IEEExplore and ACM Digital Library. The first three are recommended sources by TU Delft for the Computer Science discipline [13], which is the field of this paper. For ACM Digital Library, it was selected as it is known to be one of the most important bibliographic databases in the field of computer science [14].

2.3 Search Strategy

To make queries for the different search engines, an intersection of different concepts was used. Papers that have mention of *Facial*, *Affect*, *Recognition*, *Database* and *Rater* were targeted. These topics should at least be mentioned in the abstract or title of the paper to ensure that it is the intended topic, with the exception of the *rater* concept, which was conducted on a full-text search, explanations are given in section 2.3.1. Using table 2, the queries were built as conjunctions of disjunctions of the keywords in each column, note that the * symbols are query wildcards.

Table 1: Inclusion and Exclusion Criteria

	Criteria	**Motivation**	**Explicit Attributes**
Inclusion Criteria			
I-1	Mentions a facial affect recognition database	To focus on the core topic of facial recognition databases.	Describes datasets with explicit affect annotations in facial images or videos, i.e., facial affect recognition annotations.
I-2	Introduces a novel dataset or set of annotations	To avoid analyzing a dataset or set of annotations twice.	- Introduces a previously unpublished dataset. - Or provides new affective annotations to existing datasets of facial images or videos.
Exclusion Criteria			
E-1	The paper is not written in English	Out of the scope due to language constraints.	Papers published in languages other than English.
E-2	The paper is a review or a survey	To ensure the survey relies on primary sources.	Papers summarizing existing research.
E-3	The paper was released after April 2024 included	To limit the scope and exclude papers published after the review initiation date.	Papers with a publication date before April 2024.

Table 2: Keywords used in search query for each concept

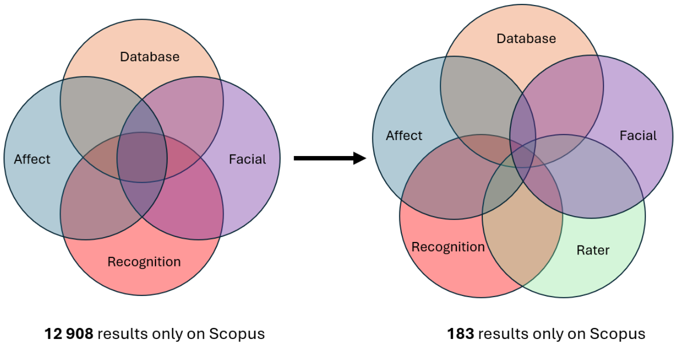
Facial	**Affect**	**Recognition**	**Database**	**Rater**
facial face* vision-based	emotion* affect* mood* feeling* facial expression	detection* recognition* prediction*	database dataset	rater* inter-rat* interrat* multiple annotators multiple annotations human annotators human annotation human-annotated human-rated human raters

2.3.1 Feasibility Criteria

To complete this project within 9 weeks, the queries were adapted for feasibility. Originally, the concept of *rater* was not included, but nearly 13,000 results were retrieved from Scopus alone. To manage this volume, the query was refined to include only papers that mention the term *rater* at least once. This adjustment significantly reduced the number of results from nearly 13,000 to less than 200 as visually shown in figure 1.

With inclusion criteria I-2, papers can be expected to include an introduction of their created database or set of annotations where the concept of *rater* is mentioned. This feasibility adjustment ensures that we identify papers discussing the *rater* concept in their full text, making it more realistic than just looking for the concept in the abstract or the title. For more details on the query development, refer to appendix A.

Figure 1: Venn Diagrams and Data Sizes Representation



2.4 Selection Process

To select papers according to the eligibility criteria and search strategy, the steps are:

1. On the different search engines, build adapted queries that match the keywords in table 2.
2. Gather all non-duplicates and start the screening process:
   (a) Screen the title and abstract, and decide whether they meet the eligibility criteria and keep, do not and reject or do not have enough information in the title and abstract and keep.
   (b) Screen the full text of the non-rejected papers and either include or exclude it from the final paper selection. During this phase, data extraction is also done for efficiency purposes.

2.5 Search Results

To report the search results, the PRISMA Flow Diagram is of help as it visually reports the different decisions made at each step. The methodology chosen is explained in the above sub-sections, and in figure 2, the exact numbers of handled papers is outputted.

From the 47 papers read, relevant data to answer the research question was gathered. An excel table was made to efficiently keep track of the different retrieved data points about each paper and its proprietary dataset. The relevant data points are shown in table 3.

Figure 2: Adapted PRISMA Flow Diagram

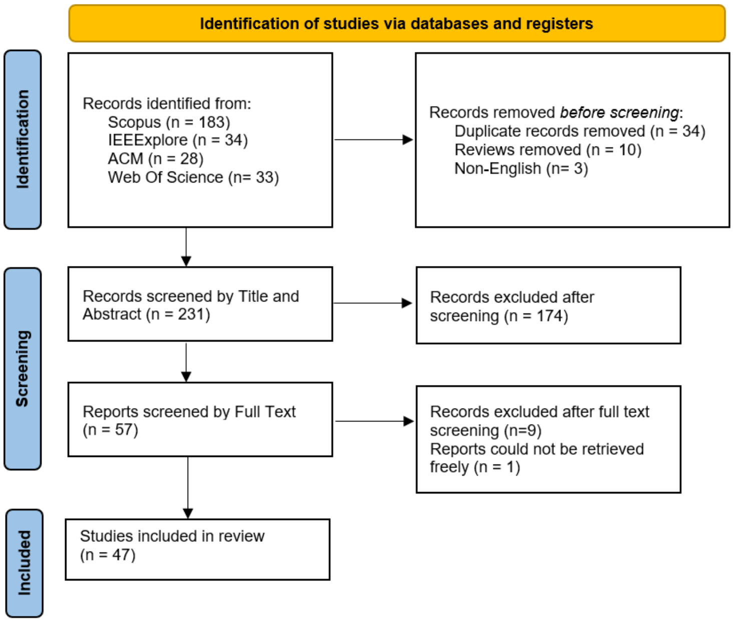

Identification of studies via databases and registers
Identification
Records identified from:
Scopus (n = 183)
IEEExplore (n = 34)
ACM (n = 28)
Web Of Science (n= 33)
Records removed *before screening*:
Duplicate records removed (n = 34)
Reviews removed (n = 10)
Non-English (n= 3)
Screening
Records screened by Title and Abstract (n = 231)
Records excluded after screening (n = 174)
Reports screened by Full Text (n = 57)
Records excluded after full text screening (n=9)
Reports could not be retrieved freely (n = 1)
Included
Studies included in review (n = 47)

Table 3: Data points for each SQ

Data	SQs
Types of affective states	1
Affect representation schemes	2, 5
Number of rating per record	3
Do they measure interrater agreement?	3a, 4
How is the interrater agreement measured?	3b, 4, 5
What is the level of agreement?	3b, 5, 6
Are there strategies in place to facilitate the interrater agreement?	3c, 5
Year of publication	4

3 Results

Once the papers gathered according to the methodology described in section 2, the relevant data to answer our research question and sub-questions are extracted. Section 3.1 is about the different affect representation schemes and affect states, then in the following section 3.2, data related to the interrater agreement in the different datasets is shown. An overall of the main strategies to facilitate IRA can be found in section 3.3 and section 3.5 is about the relationship between ARS and IRA.

3.1 Affect Representation Schemes (ARS) and Affect States

Before starting to analyse the interrater agreement variables, it is interesting to see what kind of affective states are mentioned in the different retrieved papers. It was noticed that most papers make use of Ekman's basic emotions. The choice of those 6 basic emotions have been first justified in 1971 [15], and already in 1969 [16], Ekman's paper contributed to the foundational ideas that led to the development of the Facial Action Coding System (FACS). The influence and validity of his work and collaborations can be seen in the results of this current SLR. Table 4 shows the data retrieved more extensively.

Table 4: Use of ARS per Corresponding Papers

Affect Representation Schemes	Papers	Number of Papers
Valence or arousal or dominance	[17], [18], [19], [20], [21], [22], [23], [24], [25]	9
Ekman's basic emotions (happiness, sadness, anger, fear, disgust, surprise)	[19], [20], [21], [22], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41]	22
Extended emotions (including contempt, anxious, etc.)	[18], [21], [22], [25], [39], [40], [42], [43], [44]	9
Pain, no-pain	[45]	1
Hate, no-hate	[44]	1
Smiling, frowning	[19], [32], [46], [47]	4
Mikels 8 emotions (including excitement, anger, disgust, fear, sadness)	[24], [30], [36]	3
Other complex or mixed emotions (such as amusement, awe, boredom, confusion)	[30], [33], [35], [37], [40], [41], [48], [49], [50], [51], [52], [53], [54], [55], [56]	15
Neutral	[18], [24], [25], [26], [27], [30], [32], [36], [38], [42], [33], [43], [47], [48], [50], [51], [54], [55], [56], [57], [58], [59], [60]	23

It can also be concluded by table 4 what the different ARS are. The reasons for the dataset creations are going from wanting to assess emotions of individuals in a specific situation, like while driving [34] or to create a database for a specific ethnicity [30]. The other ARS used is the *valence-arousal-dominance* one that is a dimensional ARS type, with table 5, it can be seen that only 3 papers were developing a dataset that only assessed the facial expression using a dimensional ARS type, and when used, it was mostly combined with a categorical rating as well.

Doing reversed engineering, all the ARS that include Ekman's emotions, its extended version, Mikels emotions or other complex types of emotion, involve the affective state called *emotion*. Following the logic set by Scherer [61] and Frijda [62], mood can be of low intensity but also can last over hours and days, they are more diffuse. In the datasets found, they captured faces of people reacting to a stimulus, a known trigger, that changed their emotion for a few seconds (or minutes for pain as the experiment was supervised [45]), therefore, none of them are capturing mood. Other than *emotion*, there is still a small amount of papers (6/47) that presented a dataset with alternatives of the *emotion and valence-arousal-dominance* affective state, it can be further observed with the data presented in table 5, showing the different ARS types used.

Some datasets were interested in analysing the facial affect using specific knowledge in the field. Using a comprehensive tool for measuring and describing the facial movements, known as FACS. It analyses the facial expressions with *Action Units* (AUs) and two papers took the initiative to create datasets that were measuring the different aspects of emotional expressiveness using AUs [29][44].

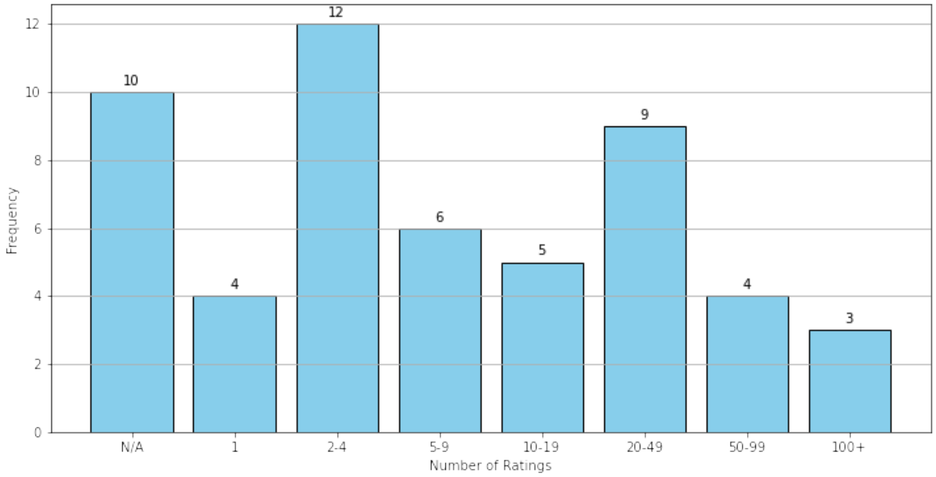
Table 5: Aggregated ARS types and Corresponding Papers

ARS types	**Papers**	**Number of Papers**
Categorical only	[26], [27], [28], [29], [30], [31], [32], [34], [35], [36], [37], [38], [39], [40], [41], [42], [33], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [63], [64], [65]	37
Dimensional only	[17], [22], [23]	3
Dimensional, Categorical	[18], [19], [20], [21], [24], [25]	6

3.2 Interrater Agreement

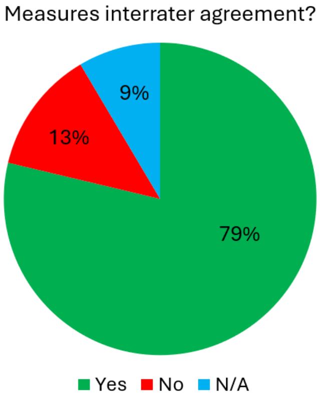
Some statistics about how interrater agreement is measured can help towards answering the final research question, in this section, SQ3, SQ3a and SQ3b are answered. As SQ3 was formulated, there is first the need to look at if multiple ratings are measured for the different records in the datasets. In figure 3, we can assess that most of them have at least 2 ratings. Note that this graph does not just have one data point per paper, some datasets were using, for instance, 1 rating per record but 8% of the data had two ratings [18]. Unfortunately, some datasets and their corresponding papers did not mention how many raters annotated each record they have. Nevertheless, with the data available, most of them are well-prepared to also measure interrater agreement and allow this paper to look further into how it is measured.

Figure 3: Distribution of Number of Ratings per Record including N/A data



Prior to observing which measurements are used for interrater agreement, figure 4 shows that indeed most papers measured interrater agreement before publishing their datasets. Some datasets use a continuous rating scale and use the average of the different ratings as the final ground truth instead of measuring the IRA [36], [57], but some are simply not giving any hint that IRA was ever measured [63], [59], [52].

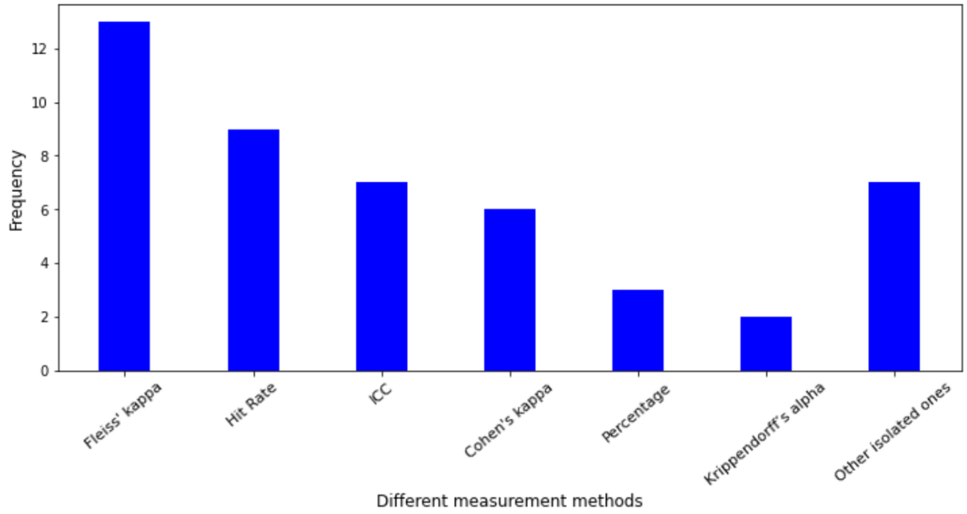
Figure 4: Pie chart showing proportion of papers measuring interrater agreement



For the proportion that measures the interrater agreement (79%), different measures were used and combined. Figure 5 visualises the frequency of each method. It is directly linked to our research sub-question (SQ3b) about the differences in interrater agreement measurement methodologies among published datasets for automatic facial affect prediction.

Fleiss' kappa was the most popular one, and it is similar to *Cohen's kappa* but at another scale, Fleiss' kappa measurement works for any number of rater [66], while Cohen's kappa only works for two raters [67]. Nevertheless, it was noticed that even when there were more than two raters' agreement to assess, it was sometimes chosen to look at Cohen's kappa to assess pair by pair inter-agreement [35] [56]. The difference between *hit rate* and *percentage* in these measurements lies in the fact that when measuring hit rate, the targeted affect is already known, or actors were asked to do a certain affect, and the accuracy between raters is measured, while percentage uses the interrater agreement value to decide upon the correct affect if the percentage is more than a threshold they decided upon. *ICC*, the intra-class correlation coefficient, measures the interrater agreement for each emotion or valence-arousal-dominance class.

Figure 5: Frequency of different interrater measurement methodologies



The results of the IRA using a kappa were assessed using the same principle described in the paper from M.D. Montefalcon, J.R. Padilla, et al [65], the measurements are similar to a normal distribution, with the peak being *substantial agreement.* Therefore, there were also some that published their datasets even though the interrater agreement of some emotions is seen as *slight agreement.* The best performed affect interrater agreement, is in the case of happiness (when part of its ARS) but fear, digust and surprise are generally part of the less agreed upon emotions. For Hit Rate, mostly all were above 70% or near, showing a good IRA, same for ICC, and there were some outstanding results, reaching 0.98 or 0.99 for some classes [23], [24], [25].

3.3 Strategies to facilitate IRA

Throughout the process of accomplishing this SLR, many papers mentioned techniques they used to ensure optimal results in their ground truth. These strategies to facilitate interrater agreement can be divided into two main categories: improving the selection and training of raters, and improving the selection of faces to be rated.

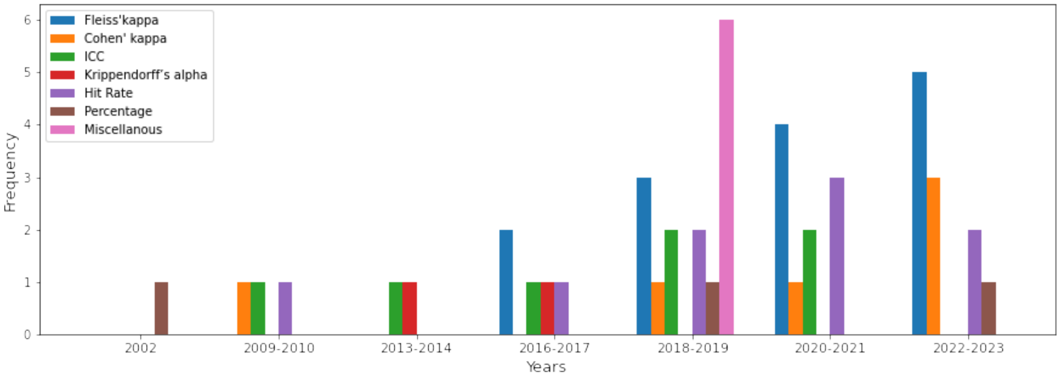
Training the raters involved training them on previous published dataset, or teaching them about FACS before they rate the provided pictures. The selection of raters was also used as a strategy, if the raters have a history of rating (crowd sourcing), then there was some threshold into their experience and performance. In 14 datasets, some or all raters were selected because they were experts with degrees in psychology, human sciences, acting, or were already familiar with FACS. Sometimes, raters were judged by their intra-rater consistency level, where a set of images were repeated randomly, to assess if the rater was staying consistent and putting enough care in the rating process. One strategy only used once was to do an empathy quotient test to verify that the raters are sensitive to emotions [31]. For the datasets that were specifically taking images of people's faces of a certain ethnicity, then it was also decided that the raters should have the same ethnicity to ensure the best possible rating.

Improving the faces' selection was done by asking actors to mimic the facial expressions. Some demanded no use of cosmetics [40] or also no glasses, hair or jewelry [21] to avoid any obstruction of the face.

3.4 IRA Methodologies over time

The papers obtained through this SLR are ranging from 2002 to 2023, allowing for a nice overview of what is happening at the beginning of the 21st century. Figure 6 shows how each methodology progressed in time over the final set of papers.

Figure 6: Frequency of different interrater measurement methodologies over time



It can be seen that the Fleiss' kappa measurement methodology (blue on the graph) has increased its popularity over time. Cohen's kappa (orange on the graph) also has been more used in 2022-2023. ICC has had a quite consistent presence over the years besides for the last year range.

3.5 ARS and IRA relationship

This section tackles SQ5, about assessing if there might be a relationship between the affect representation scheme used by datasets and their interrater agreement. This paper is not focused on defining the different ARS, and we sorted out in section 3.1, which dataset was approximately using which scheme to avoid ending up with more than a dozen of categories. Some datasets measuring the interrater agreement of dimensional ARS type rating shared the values explicitly. In the available results, some have outstanding ICC performances that were already noticed in section 3.2 [23], [24], [25], but there is also a case of only substantial agreement results [22]. The available data does not allow to confidently draw correlation conclusions, and the lack of unanimous outstanding results further complicates this. As the use of Ekman's basic emotions was popular, see table 4, a deeper look was taken into the IRA values, but no pattern could be found. An observation made thanks to a paper [50] is that if the rater is an expert or a normal individual, the results can significantly differ (90% accuracy for experts vs 59% for the normal individuals). One nearly perfect result was from an ARS including complex emotions [48], but then some others showed substantial [35] or fair agreement [40], meaning that there is no clear correlation in between the use of complex emotions and IRA.

4 Responsible Research

Responsible research practices are essential in ensuring the integrity and ethical soundness of scientific investigations. This section addresses both the methodological rigor in 4.1 and ethical considerations inherent in our review in 4.2.

4.1 Reflection upon the methodology

Conducting a systematic literature review for this survey ensures reproducibility and additionally, every step is reported according to the PRISMA guidelines [12]. The survey is conducted by one bachelor student with limited prior research experience, which may introduce potential errors, particularly during the screening phase. To mitigate this, detailed documentation and adherence to the established protocol are maintained, yet the possibility of human error or bias remains an important consideration.

4.2 Ethical viewpoint on facial affect prediction

Affect prediction, while advancing technological capabilities, raises significant ethical concerns. One major issue is the potential for misuse in surveillance or manipulation [68], where individuals' emotional states could be monitored without consent or used to influence behavior covertly. In this study, no affect prediction system is analyzed but the datasets that can be used for the systems are retrieved. All papers have disclosed their data collection techniques as well as how they recruited their participants to record their faces.

5 Discussion

In section 3, all the needed data was presented to support answering the research question. In this section, some discussions about results and practices noticed are done.

The different ARS and affect states data extracted from the papers did not allow to show a specific correlation with the final IRA of the dataset. Extracting the data just highlighted the different practices in the collected set of papers, but no other conclusions can be made.

Most papers discussed how the IRA is measured before publishing their datasets (see figure 4), and the use of Fleiss' and Cohen's kappa are recently dominant as a measurement tool (see figure 6). Even though the Fleiss' kappa and the Hit Rate methodologies have been the most used overall (see figure 5), this could show that lately, more complex methodologies are also favored as they account for chance agreement and provide a more nuanced view of interrater reliability [69] compared to Hit Rate that has a more simple and binary approach.

Interestingly, none of the reviewed papers reported unacceptably low average IRA values. This may suggest either under-reporting of low IRA cases or non-publication of datasets with inadequate reliability. It is important to note that while the average IRA is usually acceptable, the range can vary significantly, with some values being very low. For example, in the MUMBAI paper [35], the average IRA, measured using Fleiss' Kappa, was 0.381, with scores ranging from 0.013 to 0.659. To enhance annotation scores, researchers calculated the IRA between each pair of annotators using Cohen's Kappa. They then selected the annotator pair with the highest agreement for each video. This process increased the average IRA to 0.573, with scores ranging from 0.289 to 0.859. This method illustrates how parameters can be adjusted to achieve an acceptable average IRA, potentially masking the variability and low agreement. This type of strategy, rather than being genuine improvements to the IRA, can be seen as methods to present more favorable numbers and increase the dataset's validity to the public's eyes. It is crucial to be aware of this case as some papers might not disclose these practices, which can lead to an inflated perception of reliability and influence the data interpretation.

As of strategies to facilitate IRA, the ones involving filtering the faces' selection for the dataset discussed in section 3.3 might indeed improve the IRA, but then, there is a doubt those datasets will be good enough for the machine learning algorithm training of real world applications. The data will only contain the good scenarios with the perfect facial expression and lightning conditions, but there will be no adapted training available for real-world scenario data.

A limitation this SLR encountered was time. If more time and man-power could have been allocated for this 9-week project, the feasibility constraints could have been revisited to allow more papers to be included and have a larger dataset of datasets by the end. This could have allowed to make stronger statistical analyses, and potentially find correlations in the practices surrounding the use of IRA.

6 Conclusions and Future Work

The aim of this systematic literature review was to highlight the differences in interrater agreement methodologies among the published datasets for automatic facial affect prediction. Therefore, the question was divided into sub-questions to be able to identify the multiple facets of the datasets and their methodologies.

The review revealed the different affect states targeted and affect representation schemes used, but no direct correlation was found between the measures employed and the resulting IRA. The analysis of the different ways to measure interrater agreement showed that Fleiss' kappa and Cohen's kappa were prominently used throughout the datasets that measured the interrater agreement. Additionally, strategies to facilitate IRA were identified as well as some potential methods used by authors to present IRA results in a more favorable light, which could mislead readers.

This SLR can be a basis for a future work linking IRA with the empirical performances of the systems using these databases. Investigating the relationship between IRA and empirical performance could shed light on whether the variables under study are adequately represented and if they contribute to improving the robustness of affective computing systems in practical applications.

References

[1] Y. Li and W. Zhang, “Emotional Recognition of Human-Computer Interaction Technology: Automatic Driving Safety,” in *2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE)*, pp. 199–202, Aug. 2023.

[2] F. J. Ming, S. Shabana Anhum, S. Islam, and K. H. Keoy, “Facial Emotion Recognition System for Mental Stress Detection among University Students,” in *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pp. 1–6, July 2023.

[3] J. A. Russell, “Core affect and the psychological construction of emotion,” *Psychological Review*, pp. 145–172, 2003.

[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] R. A. Calvo and S. D’Mello, “Affect detection: An interdisciplinary review of models, methods, and their applications,” *IEEE Transactions on Affective Computing*, vol. 1, pp. 17–28, 2010.

[6] I. Siegert, R. Bock, and A. Wendemuth, “Inter-rater reliability for emotion annotation in human-computer interaction – comparison and methodological improvements,” *Journal of Multimodal User Interfaces*, vol. 8, pp. 17–28, 01 2014.

[7] M. L. McHugh, “Interrater reliability: the kappa statistic,” *Biochem Med (Zagreb)*, vol. 22, p. 276â282, 2012.

[8] I. Siegert, R. Bock, and A. Wendemuth, “Inter-rater reliability for emotion annotation in humanâcomputer interaction: comparison and methodological improvements,” *Journal on Multimodal User Interfaces*, vol. 8, pp. 17–28, 2014.

[9] C. Cooper, A. Booth, and J. Varley-Campbell, “Defining the process to literature searching in systematic reviews: a literature review of guidance and supporting studies.,” *BMC Med Res Methodol*, vol. 8, pp. 145–172, 2018.

[10] C. Sohrabi, T. Franchi, G. Mathew, A. Kerwan, M. Nicola, M. Griffin, M. Agha, and R. Agha, “PRISMA 2020 statement: What’s new and the importance of reporting guidelines,” *International Journal of Surgery*, vol. 88, p. 105918, Apr. 2021.

[11] M. J. Grant and A. Booth, “A typology of reviews: an analysis of 14 review types and associated methodologies,” *Health Information Libraries Journal*, vol. 26, pp. 91–108, 06 2009.

[12] "Prisma 2020 checklist." https://www.prisma-statement.org/s/PRISMA_2020_checklist-fxke.docx [Accessed: April 2024)].

[13] "Tu delft library databases." https://databases.tudl.tudelft.nl/?f=EEMCS&d=CS&t=&q=&y0=research%20data&y1=reference&y2=reports&y3=articles&y4=&y5=standards&y6=e-books&y7=patent%20information&y8=statistics&y9=educational%20resource&y10=theses&y11=e-journals&y12=catalogue [Accessed: April 2024)].

[14] "Acm digital library - an archive of original research." https://www.acm.org/publications/digital-library [Accessed: April 2024)].

[15] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," in *Journal of Personality and Social Psychology*, vol. 17, pp. 124–129, 1971.

[16] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Semiotica*, vol. 1, no. 1, pp. 49–98, 1969.

[17] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "Afew-va database for valence and arousal estimation in-the-wild," *Image and Vision Computing*, vol. 65, pp. 23–36, 2017.

[18] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2019.

[19] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The belfast induced natural emotion database," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 32–41, 2012.

[20] S. M. Kim, Y. J. Kwon, S. Y. Jung, M. J. Kim, Y. S. Cho, H. T. Kim, K. C. Nam, H. J. Kim, K. H. Choi, and J. S. Choi, "Development of the korean facial emotion stimuli: Korea university facial expression collection 2nd edition," *FRONTIERS IN PSYCHOLOGY*, vol. 8, 2017.

[21] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[22] R. Gupta, M. K. Abadia, J. A. C. Cabre, F. Morreale, T. H. Falk, and N. Sebe, "A quality adaptive multimodal affect recognition system for user-centric multimedia indexing," p. 317â320, 2016.

[23] T. M. Sutton, A. M. Herbert, and D. Q. Clark, "Valence, arousal, and dominance ratings for facial stimuli," *Quarterly Journal of Experimental Psychology*, vol. 72, no. 8, pp. 2046–2055, 2019.

[24] I. A. M. Verpaalen, G. Bijsterbosch, L. Mobach, G. Bijlstra, M. Rinck, and A. M. Klein, "Validating the radboud faces database from a childâs perspective," *Cognition and Emotion*, vol. 33, no. 8, pp. 1531–1547, 2019.

[25] G. Bijsterbosch, L. Mobach, I. A. M. Verpaalen, G. Bijlstra, J. L. Hudson, M. Rinck, and A. M. Klein, "Validation of the child models of the radboud faces database by children," *INTERNATIONAL JOURNAL OF BEHAVIORAL DEVELOPMENT*, vol. 45, no. 2, pp. 146–152, 2021.

[26] J. Tejada, R. M. K. Freitag, B. F. M. Pinheiro, P. B. Cardoso, V. R. A. Souza, and L. S. Silva, "Building and validation of a set of facial expression images to detect emotions: a transcultural study," *Psychological Research*, vol. 86, no. 6, pp. 1996–2006, 2022.

[27] R. Vemulapalli and A. Agarwala, "A compact embedding for facial expression similarity," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 5676–5685.

[28] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu, "Dfew: A large-scale database for recognizing dynamic facial expressions in the wild," in *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2881–2889.

[29] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[30] Y. Z. Tu, D. W. Lin, A. Suzuki, and J. O. S. Goh, "East asian young and older adult perceptions of emotional faces from an age- and sex-fair east asian facial expression database," *Frontiers in Psychology*, vol. 9, no. NOV, 2018.

[31] J. Yang, Q. Huang, T. Ding, D. Lischinski, D. Cohen-Or, and H. Huang, "Emoset: A large-scale visual emotion dataset with rich attributes," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20326–20337.

[32] C. C. Chen, S. L. Cho, K. Horszowska, M. Y. Chen, C. C. Wu, H. C. Chen, Y. Y. Yeh, and C. M. Cheng, "A facial expression image database and norm for asian population: A preliminary report," in *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 7242.

[33] D. Y. Liliana, T. Basaruddin, and I. I. D. Oriza, "The indonesian mixed emotion dataset (imed): A facial expression dataset for mixed emotion recognition."

[34] M. Weber, J. Giacomin, A. Malizia, L. Skrypchuk, V. Gkatzidou, and A. Mouzakitis, "Investigation of the dependency of the driversâ emotional experience on different road types and driving conditions," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 65, pp. 107–120, 2019.

[35] M. Doyran, A. Schimmel, P. Baki, K. Ergin, B. TÃŒrkmen, A. A. Salah, S. C. J. Bakkes, H. Kaya, R. Poppe, and A. A. Salah, "Mumbai: multi-person, multimodal board game affect and interaction analysis dataset," *Journal on Multimodal User Interfaces*, vol. 15, no. 4, pp. 373–391, 2021.

[36] M. C. Voelkle, N. C. Ebner, U. Lindenberger, and M. Riediger, "A note on age differences in mood-congruent vs. mood-incongruent emotion processing in faces," *Frontiers in Psychology*, vol. 5, no. JUN, 2014.

[37] A. Miolla, M. Cardaioli, and C. Scarpazza, "Padova emotional dataset of facial expressions (pedfe): A unique dataset of genuine and posed emotional facial expressions," *BEHAVIOR RESEARCH METHODS*, vol. 55, no. 5, pp. 2559–2574, 2023.

[38] E. Lyakso, O. Frolova, A. Nikolaev, E. Kleshnev, P. Grave, A. Ilyas, O. Makhnytkina, R. Nersisson, A. Mary Mekala, and M. Varalakshmi, "Recognition of the emotional state of children by video and audio modalities by indian and russian experts," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14338 LNAI, pp. 469–482.

[39] J. M. Girard, W. S. Chu, L. A. Jeni, J. F. Cohn, F. De La Torre, and M. A. Sayette, "Sayette group formation task (gft) spontaneous facial expression database," in *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heterogeneous Face Recognition, HFR 2017, Joint Challenge on Dominant and Complementary Emotion Recognition Using Micro Emotion Features and Head-Pose Estimation, DCER and HPE 2017 and 3rd Facial Expression Recognition and Analysis Challenge, FERA 2017*, pp. 581–588.

[40] P. Saha, D. Bhattacharjee, B. K. De, and M. Nasipuri, "A thermal blended facial expression analysis and recognition system using deformed thermal facial areas," *International Journal of Image and Graphics*, vol. 22, no. 5, 2022.

[41] L. Teijeiro-Mosquera, J. I. Biel, J. L. Alba-Castro, and D. Gatica-Perez, "What your face vlogs about: Expressions of emotion and big-five traits impressions in youtube," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 193–205, 2015.

[42] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[43] R. C. Gur, R. Sara, M. Hagendoorn, O. Marom, P. Hughett, L. Macy, T. Turner, R. Bajcsy, A. Posner, and R. E. Gur, "A method for obtaining 3-dimensional facial expressions and its standardization for use in neurocognitive studies," *JOURNAL OF NEUROSCIENCE METHODS*, vol. 115, no. 2, pp. 137–143, 2002.

[44] V. Lin, J. M. Girard, M. A. Sayette, and L. P. Morency, "Toward multimodal modeling of emotional expressiveness," in *ICMI 2020 - Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 548–557.

[45] M. S. H. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Y. Meng, A. Kemp, M. Shafizadeh, A. C. Elkins, N. Kanakam, A. de Rothschild, N. Tyler, P. J. Watson, A. C. D. Williams, M. Pantic, and N. Bianchi-Berthouze, "The automatic detection of chronic pain-related expression: Requirements, challenges and the multimodal <i>emopain</i> dataset," *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, vol. 7, no. 4, pp. 435–451, 2016.

[46] C. Tang, W. Zheng, Y. Zong, Z. Cui, N. Qiu, and S. Yan, "Automatic smile detection of infants in mother-infant interaction via cnn-based feature learning," 2018.

[47] F. Heydari, S. Sheybani, and A. Yoonessi, “Iranian emotional face database: Acquisition and validation of a stimulus set of basic facial expressions,” *Behavior Research Methods*, vol. 55, no. 1, pp. 143–150, 2023.

[48] C. Bian, Y. Zhang, D. Wang, Y. Liang, B. Wu, and W. Lu, “An academic emotion database and the baseline evaluation,” in *13th International Conference on Computer Science and Education, ICCSE 2018*, pp. 378–383.

[49] M. S. Benda and K. S. Scherf, “The complex emotion expression database: A validated stimulus set of trained actors,” *PLoS ONE*, vol. 15, no. 2, 2020.

[50] J. Ma, B. Yang, R. Luo, and X. Ding, “Development of a facial-expression database of chinese han, hui and tibetan people,” *International Journal of Psychology*, vol. 55, no. 3, pp. 456–464, 2020.

[51] B. Novello, A. Renner, G. Maurer, S. Musse, and A. Arteche, “Development of the youth emotion picture set,” *Perception*, vol. 47, no. 10-11, pp. 1029–1042, 2018.

[52] S. L. Happy, P. Patnaik, A. Routray, and R. Guha, “The indian spontaneous expression database for emotion recognition,” *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 131–142, 2017.

[53] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias, “Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4451 LNAI, pp. 91–112.

[54] B. De Carolis, F. DâErrico, N. Macchiarulo, M. Paciello, and G. Palestra, “Recognizing cognitive emotions in e-learning environment,” in *Communications in Computer and Information Science*, vol. 1344, pp. 17–27.

[55] Y. Xu, Y. Li, Y. Chen, H. Bao, and Y. Zheng, “Spontaneous visual database for detecting learning-centered emotions during online learning,” *Image and Vision Computing*, vol. 136, 2023.

[56] M. E. Hoque, R. El Kaliouby, and R. W. Picard, “When human coders (and machines) disagree on the meaning of facial affect in spontaneous videos,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5773 LNAI, pp. 337–343.

[57] H. Liang, P. Perona, and G. Balakrishnan, “Benchmarking algorithmic bias in face recognition: An experimental approach using synthetic faces and human evaluation,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4954–4964.

[58] R. Verma, N. Kalsi, N. P. Shrivastava, and A. Sheerha, “Development and validation of the aiims facial toolbox for emotion recognition,” *Indian Journal of Psychological Medicine*, vol. 45, no. 5, pp. 471–475, 2023.

[59] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, “Emotion recognition in the wild challenge 2013,” 2013.

[60] L. Singh, N. Aggarwal, and S. Singh, "Pumave-d: panjab university multilingual audio and video facial expression dataset," *Multimedia Tools and Applications*, vol. 82, no. 7, pp. 10117–10144, 2023.

[61] K. R. Scherer, "What are emotions? and how can they be measured?," *Social Science Information*, vol. 44, pp. 695–729, 2005.

[62] N. H. Frijda, "The psychologistâs point of viewâ," *Cambridge University Press*, 1986.

[63] T. Hama and M. Koeda, "Characteristics of healthy japanese young adults with respect to recognition of facial expressions: a preliminary study," *BMC Psychology*, vol. 11, no. 1, 2023.

[64] K. M. Chung, S. Kim, W. H. Jung, and Y. Kim, "Development and validation of the yonsei face database (yface db)," *FRONTIERS IN PSYCHOLOGY*, vol. 10, 2019.

[65] M. D. Montefalcon, J. R. Padilla, J. Paulino, J. Go, R. L. Rodriguez, and J. M. Imperial, "Understanding facial expression expressing hate from online short-form videos," 2021.

[66] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[67] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[68] H. Lanlan and L. Xiaoyi, "China leads in emotion recognition tech, reinforces privacy rules to tackle abuse." `https://www.globaltimes.cn/page/202103/1217212.shtml` [Accessed: May 2024)].

[69] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: The kappa statistic," *Family medicine*, vol. 37, no. 5, pp. 360–363, 2005.

A Queries

Query development for “Facial Affect Recognition”

Scopus query:

((TITLE-ABS-KEY ("facial" OR "face*" OR "vision-based")) AND (TITLE-ABS-KEY ("emotion*" OR "affect*" OR "mood*" OR "feeling*" OR "facial expression"))) AND (TITLE-ABS-KEY ("recogni*" OR "predict*" OR "detect*")) AND (TITLE-ABS-KEY ("interrater*" OR "inter-rater*" OR "multiple raters" OR "between raters"))	173	It should be limited to facial affect recognition/detection/prediction, so no need to include all forms of the noun.
((TITLE-ABS-KEY ("facial" OR "face*" OR "vision-based")) AND (TITLE-ABS-KEY ("emotion*" OR "affect*" OR "mood*" OR "feeling*" OR "facial expression"))) AND (TITLE-ABS-KEY ("recognition*" OR "prediction*" OR "detection*")) AND (TITLE-ABS-KEY ("interrater*" OR "inter-rater*" OR "multiple raters" OR "between raters"))	91	That’s good but in one of the research sub questions, I need to identify the amount of raters, so it should not be assumed that there are multiple raters, no matter what my opinion is.
(TITLE-ABS-KEY ("facial*" OR "face*" OR "vision-based")) AND (TITLE-ABS-KEY ("emotion*" OR "affect*" OR "mood*" OR "feeling*" OR "facial expression")) AND (TITLE-ABS-KEY ("recognition*" OR "detection*" OR "prediction*")) AND (TITLE-ABS-KEY ("rater*"))	151	Using “rater*” includes more papers and does not imply that there is more than one rater, but we lose interrater and inter-rater only mentions. Doing “*rater*” seems too inclusive to other words.
((TITLE-ABS-KEY ("facial" OR "face*" OR "vision-based")) AND (TITLE-ABS-KEY ("emotion*" OR "affect*" OR "mood*" OR "feeling*" OR "facial expression"))) AND (TITLE-ABS-KEY ("recognition*" OR "prediction*" OR "detection*")) AND (TITLE-ABS-KEY ("interrater*" OR "inter-rater*" OR "multiple raters" OR "between raters" OR "rater*"))	197	That includes all types of raters, but we need to make sure that they mention database as we need to analyze it.
((TITLE-ABS-KEY ("facial" OR "face*" OR "vision-based")) AND (TITLE-ABS-KEY ("emotion*" OR "affect*" OR "mood*" OR "feeling*" OR "facial expression"))) AND (TITLE-ABS-KEY ("recognition*" OR "prediction*" OR "detection*")) AND (TITLE-ABS-KEY ("interrater*" OR "inter-rater*" OR "multiple raters" OR "between raters" OR "rater*")) AND (TITLE-ABS-KEY ("dataset" OR "database"))	32	That includes all conditions, and we get a reasonable amount of papers. But maybe we can make the rater concept broader.

((TITLE-ABS-KEY ("facial" OR "face*" OR "vision-based")) AND (TITLE-ABS-KEY ("emotion*" OR "affect*" OR "mood*" OR "feeling*" OR "facial expression"))) AND ((TITLE-ABS-KEY ("recognition*" OR "prediction*" OR "detection*")) AND (TITLE-ABS-KEY ("dataset" OR "database")))	12908	Just to try if rater is necessary: If we remove the mention of raters, then we end up with way too many results, so the previous query seems to be the right one.
((TITLE-ABS-KEY ("facial" OR "face*" OR "vision-based")) AND (TITLE-ABS-KEY ("emotion*" OR "affect*" OR "mood*" OR "feeling*" OR "facial expression"))) AND (TITLE-ABS-KEY ("recognition*" OR "prediction*" OR "detection*")) AND (TITLE-ABS-KEY (rater* OR "multiple annotators" OR "multiple annotations" OR "human annotators" OR "human annotation" OR "human-annotated" OR "human-rated" OR "human raters" OR interrat* OR inter-rat*)) AND (TITLE-ABS-KEY ("dataset" OR "database"))	58	Adding the notion of annotator and human-rated to expand the rating concept. But the concept of rater can be mentioned anywhere in the text, not only title, abstract keyword as the database might be described inside the text.
((TITLE-ABS-KEY ("facial" OR "face*" OR "vision-based")) AND (TITLE-ABS-KEY ("emotion*" OR "affect*" OR "mood*" OR "feeling*" OR "facial expression"))) AND (TITLE-ABS-KEY ("recognition*" OR "prediction*" OR "detection*")) AND (ALL (rater* OR "multiple annotators" OR "multiple annotations" OR "human annotators" OR "human annotation" OR "human-annotated" OR "human-rated" OR "human raters" OR interrat* OR inter-rat*)) AND (TITLE-ABS-KEY ("dataset" OR "database"))	**183** **(** **173 when reviews and non english excluded)**	**Concept of rater adapted.** **→ Final SCOPUS query**
((ALL ("facial" OR "face*" OR "vision-based")) AND (ALL ("emotion*" OR "affect*" OR "mood*" OR "feeling*" OR "facial expression"))) AND (ALL ("recognition*" OR "prediction*" OR "detection*")) AND (ALL (rater* OR "multiple annotators" OR "multiple annotations" OR "human annotators" OR "human annotation" OR "human-annotated" OR "human-rated" OR "human raters" OR interrat* OR inter-rat*)) AND (ALL ("dataset" OR "database"))	3137	Just to try if everything can be looked up in the whole text, but it is unfeasible for this project, we get too many results.

Final for other bibliographic resources:

ACM:

[[Abstract: "facial"] OR [Abstract: "face*"] OR [Abstract: "vision-based"]] AND [[Abstract: "emotion*"] OR [Abstract: "affect*"] OR [Abstract: "mood*"] OR [Abstract: "feeling*"] OR [Abstract: "facial expression"]] AND [[Abstract: "recognition*"] OR [Abstract: "prediction*"] OR [Abstract: "detection*"]] AND [[Abstract: "dataset"] OR [Abstract: "database"]] AND [[All: rater*] OR [All: "multiple annotators"] OR [All: "multiple annotations"] OR [All: "human annotators"] OR [All: "human annotation"] OR [All: "human-annotated"] OR [All: "human-rated"] OR [All: "human raters"] OR [All: interrat*] OR [All: inter-rat*]] → gets 28

IEEE:

(((("All Metadata":"facial" OR "All Metadata":"face*" OR "All Metadata":"vision-based") AND ("All Metadata":"emotion" OR "All Metadata":"affect*" OR "All Metadata":"mood" OR "All Metadata":"feeling*" OR "All Metadata":"facial expression") AND ("All Metadata":"recognition*" OR "All Metadata":"prediction*" OR "All Metadata":"detection*") AND ("All Metadata":"dataset" OR "All Metadata":"database") AND ("All Metadata":"rater*" OR "All Metadata":"multiple annotators" OR "All Metadata":"multiple annotations" OR "All Metadata":"human annotators" OR "All Metadata":"human annotation" OR "All Metadata":"human-annotated" OR "All Metadata":"human-rated" OR "All Metadata":"human raters" OR "All Metadata":"interrat*" OR "All Metadata":"inter-rat*")))) → gets 34

Web of Science:
(TS=("facial" OR "face*" OR "vision-based")) AND (TS=("emotion*" OR "affect*" OR "mood*" OR "feeling*" OR "facial expression")) AND (TS=("recognition*" OR "prediction*" OR "detection*")) AND (ALL=("rater*" OR "multiple annotators" OR "multiple annotations" OR "human annotators" OR "human annotation" OR "human-annotated" OR "human-rated" OR "human raters" OR "interrat*" OR "inter-rat*")) AND (TS=("dataset" OR "database")) → gets 33

Total of 231 papers to screen after removing duplicates, reviews and non-English papers.