A Systematic Review of Facial Detection and Expression Recognition in Groups of People

Felipe Zago Canal, Post Graduate Program in Information Technology and Communication (PPGTIC) - Federal University of Santa Catarina - Brazil, felipe.canal@grad.ufsc.br - ORCID: 0000-0002-9307-7546
Dennis Paz Lopez, Department of Computing (DEC) - Federal University of Santa Catarina - Brazil, ppazlopez@gmail.com - ORCID: 0009-0004-6946-9234
Dra. Eliane Pozzebon, Post Graduate Program in Information Technology and Communication (PPGTIC) - Federal University of Santa Catarina - Brazil, eliane.pozzebon@ufsc.br - ORCID: 0000-0002-4237-6589
Dr. Antonio C. Sobieranski, Post Graduate Program in Information Technology and Communication (PPGTIC) - Federal University of Santa Catarina - Brazil, a.sobieranski@ufsc.br - ORCID: 0000-0003-1147-1900

Resumo: Expression recognition from facial inputs based on scene-based group of individuals is a very important approach, presenting potential applications for business, security, education, and healthcare areas. Recently, many approaches have been proposed to address this problem, having as a general solution the formulation of the problem as an extension from the single-face detection approach, while other approaches are specifically designed taking into account the multi-face scenario. This study presents a systematic literature review on the state-of-the-art group-level expression detection and recognition technique, based on facial images. The paper has, as its main goal, the identification of the most used strategies published over the past few years to interpret and recognize facial emotion expressions in groups of people. For this purpose, a total of 319 papers were collected from multiple well-established scientific databases (ACM Digital Library, IEEE Xplore, and Scopus) and, after applying the methodology for systematic literature and its inclusion and exclusion criterion, a total of 14 papers were analyzed from the literature, totaling 16 distinct methods. The obtained analysis demonstrates an extensive application for Convolutional Neural Networks (CNNs) compared to other categories of methods. Although predominantly used, the overall scores presented by CNNs were not the best suited across the evaluated methods. This literature review suggests that besides the good results achieved, there is still an open problem and a significant range for improvements, especially for the CNN counterpart.

Palavras-chave: Facial Expression Recognition, Emotion Classification, Group-level Facial Exrpression Recognition.

1. Introduction

Information and Communication Technologies (ICTs) have changed the way humans perform daily activities in the last past decade (Văidean e Achim, 2022). Many fields have made significant progress with the application of new ICTs, thanks to the computational power capacity improvement and the newly designed machine learning techniques (Ahmad *et al.*, 2022; Sigov *et al.*, 2022).

With the evolution of ICTs taking place, education must follow the trends and make good use of these technologies to better help students and professors in the teaching and learning process. In the last few years, mainly because of the pandemic COVID-19, the education process had to be adapted. The application of Intelligent Tutoring System (ITSs) had gained space to implement computational solutions from the fields of cognitive sciences, education sciences, and artificial intelligence (Hwang, 2003; Graesser; Conley

e Olney, 2012). However, ITSs are not perfect and one of the downsides of using them to teach and learn is the lack of affection involved. Emotions such as anxiety, anger, confusion, and boredom can have a negative impact on learning (Rebolledo-Mendez *et al.*, 2022). One way to work around this problem is enabling the ITS to detect (from facial expressions, for instance) the emotion of the students and, therefore, intervene in the learning process to help them.

Performing Facial Expression Recognition computationally is still a challenging task. Besides being a very natural task for humans, it is a field of computer vision that requires great computational power and development and, therefore, is an easy problem to be solved by machines (Biswas e Sil, 2015). Nevertheless, the applications that can benefit from this kind of technology vary across many areas, such as robotics, digital marketing, and education (Lopes; Aguiar e Oliveira-Santos, 2015). Although this technology is still unresolved, over the literature, there are many works that have been able to achieve good performance and results in general. Some deep learning techniques that apply Artificial Neural Networks (ANNs) (Ali *et al.*, 2015; Lopes *et al.*, 2017; Jain; Shamsolmoali e Sehdev, 2019) and some more classic methods such Support Vector Machine (SVM) and Fuzzy logic (Ali; Iqbal e Choi, 2016; Happy e Routray, 2014; Biswas e Sil, 2015; Ghasemi e Ahmady, 2014) are some examples of techniques applied to this kind of problem.

Previous mappings of the state-of-the-art literature about Facial Emotion Recognition (FER) have been developed recently (Canal *et al.*, 2022; Li e Deng, 2020) but none of them have focused on algorithms capable of detecting and recognizing emotions in groups of people (group-level FER). The individual has always been the subject of study and, because of that, group-level FER is still in its infancy compared to individual FER (Quach *et al.*, 2022). With the advance of group-level FER, systems to perform the automatic selection of pictures for a photo album can be developed, for example. This kind of algorithm may be employed to assist social scientists and researchers in the field of education to analyze the interactions between students in the collaborative learning process (Huang *et al.*, 2019).

To address the aforementioned drawback, this review intends to focus on group-level FER algorithms. The main goal of this work is to identify the state-of-the-art in the area of facial expression/emotion recognition in groups, and terms of detected emotions, accuracy, databases, and general effectiveness. A systematic literature review was conducted in order to analyze the current research in the field, and the raised methods allowed us to suggest future directions for research.

2. Methodology

A Systematic Literature Review (SLR), according to (Kitchenham, 2004), is a secondary study as it is developed based on a group of primary studies. At the SLR, all primary studies about a determined research question, topics or phenomena of interest are identified, evaluated and interpreted. The online tool Parsifal*was used to organize, plan and conduct the research, collaborating to the systematization of the research according to (Kitchenham, 2004) in the following aspects: Research objectives, Research questions, Articles bases, and Quality criterion.

Based on the main goals of this review, the research questions are presented in Table 1.

*https://parsif.al/

Tabela 1. Research Questions.

Code	Research Questions
RQ1	What are the technologies used to detect and recognize facial expressions/emotions in groups of people?
RQ2	What are the main advantages and disadvantages of each method?
RQ3	What is the general accuracy achieved by these algorithms?

2.1. Eligibility Criterion

To be part of this work, studies containing the development of methods to detect and classify facial emotions in groups of people were selected. Studies about the usage of existing methods were also considered. To help the understanding of the objectives of this study, the PICOC (Wohlin *et al.*, 2012) strategy was defined as follows:

- **Population:** General public composed by more than one person (ex: study group);
- **Intervention:** Multiple users facial detection and expression/emotion recognition algorithms;
- **Comparison:** Do not apply;
- **Outcomes:** Mapping of state-of-the-art algorithms to detect and recognize facial expressions/emotions in groups of people;
- **Context:** Mainly any group of people which is wanted to recognize expressions/emotions (ex: classrooms, audiences...).

To establish the acceptance or rejection of articles, according tho the research methodology applied in this study, inclusion and exclusion criteria were employed, as presented in Table 2.

2.2. Articles research

To cover both national and international scenarios, the Scopus[†] database was adopted as one of the research's sources of papers to be considered for this SLR. Besides Scopus, the search for articles was executed based on the open access articles contained in the IEEE [‡]and ACM [§]databases, according to all inclusion criterion (Table 2). The main terms selected were: *facial expression recognition*, *facial emotion recognition*, and *group*. The search was restricted to the years between 2018 and 2022 because with the rapid changes in technology and the face recognition applications, the focus is in the latest advances and challenges. The reason for this choice is because this review aims at addressing the state-of-the-art in the field; hence, we focus on recent researches. With these requirements, the search strings were defined as presented in Table 3 along with the results retrieved from each one of the databases.

The execution of this string in the database resulted in 319 articles that were further analyzed, as described in the following sections.

2.3. Studies selection

Once the studies were recovered from the databases, a selection process was conducted to evaluate each article. Firstly, duplicated articles, studies without abstracts, and short documents such as abstracts or expanded abstracts were removed from the

[†]https://www.scopus.com/
[‡]https://ieeexplore.ieee.org
[§]https://dl.acm.org

Tabela 2. Inclusion and Exclusion Criterion.

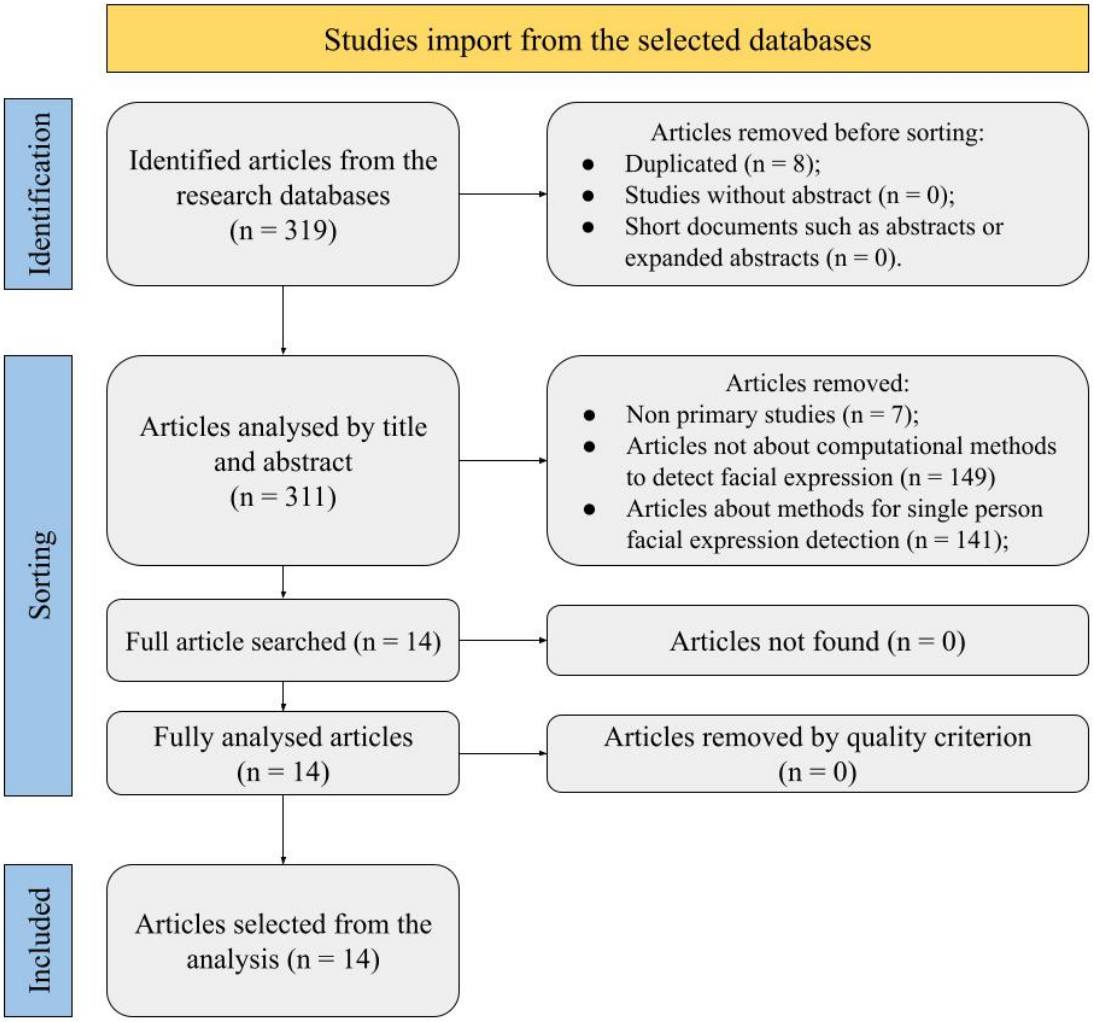
Inclusion Criterion	
Code	**Description**
I1	Articles in English or Portuguese.
I2	Articles published between 2018 and 2022.
I3	Articles that describe the development and/or usage of a method of facial expressions/ emotions recognition in a group of people.
I4	Articles that present results in terms of accuracy.
Exclusion Criterion	
Code	**Description**
E1	Short articles such as abstracts or extended/expanded abstracts.
E2	Articles without abstract.'
E3	Non-primary studies.
E4	Articles about methods for single-person facial expression detection.
E5	Articles not about computational methods to recognize facial expression.

Tabela 3. Search queries and results for each database.

Database	**Query**	**Results**
Scopus	TITLE-ABS ((facial AND (expression OR emotion) AND recognition) AND group) AND (LIMIT-TO (OA , "all")) AND (LIMIT-TO (PUBSTAGE , "final")) AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (PUBYEAR , 2022) OR LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018)) AND (LIMIT-TO (LANGUAGE , "English") OR LIMIT-TO (LANGUAGE , "Portuguese"))	266
IEEE	(("All Metadata":facial) AND (("All Metadata":expression) OR ("All Metadata":emotion)) AND ("All Metadata":recognition) AND ("All Metadata":group))	18
ACM	(Title:(facial) OR Abstract:(facial)) AND (Title:(expression) OR Title:(emotion) OR Abstract:(expression) OR Abstract:(emotion)) AND (Title:(recognition) OR Abstract:(recognition)) AND (Title:(group) OR Abstract:(group))	35

review, resulting in 311 pieces. In the next step, the title and abstract of every work were investigated to select only the relevant material for this work. In this step, 7 papers were excluded for not being primary studies, such as scope and systematic reviews; 149 studies were excluded because they were not really about computational methods of facial expression detection; and 141 were excluded from the review for being about methods to detect facial expressions in individual person images. This step resulted in 14 papers that were subsequently analyzed at their full length. We were able to access all 14 papers and none of them was excluded by the quality criterion. Therefore, 14 papers composed this review as the result of the study selection process. The complete process is shown in Figure 1.

Figura 1. Identification, sorting and included papers diagram



After the initial analysis, the resulting works were evaluated in their entirety to find the answers to the quality assessment questions. To eliminate low-quality or low-interest works from this review, the quality assessment was defined as follows:

- **AQ1:** Does the article clearly describe the method developed or applied?
- **AQ2:** Does the article present the database(s) used for training?
- **AQ3:** Can the method in the article be replicated with the information presented?
- **AQ4:** Was the method tested properly?

To evaluate the articles, three levels of response were used, each one with a specific grade (Yes: 1.0; Partially: 0.5; No: 0.0). Therefore, papers evaluated with

less than 2.0 points should be excluded from the review. Since all the analysed articles achieved 2.0 point or more, none were excluded by the quality criterion.

3. Results

As shown in Figure 1, from the total of 319 articles recovered from the databases, 8 were duplicated, 7 were non-primary studies, and 290 here not relevant to this study either because they were about facial expression recognition applied to a single person (not group-level) or not about computational methods of facial expression recognition at all. Therefore, the remaining 14 articles that were evaluated as relevant for this study were read in their entirety and the quality assessment questions were answered for all of them. The paper's evaluations are presented in Table 4. As described in section 2.3, the papers that were not able to achieve a minimum score of 2.0 should be excluded from the analysis, therefore, none of the selected articles were excluded by having a low quality according to the quality assessment proposed.

Although the sample size is relatively small (14 papers), in Table 4 it is possible to notice that China, India, and USA have the most quantity of articles published about the specific topic of group-level FER. It is also possible to identify that the majority of the articles selected range in publication from 2018 to 2020. It is interesting how the distribution of publications by year is similar between 2018 and 2020 but drastically reduces in the subsequent years. This may be due to the fact that the pandemic has had a significant impact on the ability to produce studies of this nature or simply because, in this scenario, another kind of research was more worth publishing.

The remaining papers were analysed according to method applied in their work. This may not be the most fair way of comparing multiple results specially in the computer vision problems where the dataset used, for example, is of great impact to the results. However, the method was analysed with the purpose of get this information organized as it is the most clear and easy understandable way.

For better understanding and to simplify the abbreviations used in the adjacent sections of this work, the following list mention all the algorithms mapped through the papers and their abbreviation:

- AM: Attention Mechanisms
- BFN: Big Face Network
- CNN: Convolutional Neural Network
- GF: Geometric Features
- LSTM: Long Short-Term Memory
- MTCNN: Multi-Task Cascaded Convolutional Neural Network
- MFM: Multiscale Feature Map
- NVPF: Non-Volume Preserving Fusion
- QLZM: Quantised Local Zernike Moments
- RNN: Recurrent Neural Network
- RF: Random Forest
- SFN: Small Face Network
- SVM: Support Vector Machine
- SVR: Support Vector Regression
- TNVPF: Network Temporal Non-volume Preserving Fusion
- TSM: Temporal Shift Module
- VJ: Viola-Jones

From the 14 papers considered, only two of them based their approach on geometric features from facial landmarks (Palestra e Pino, 2020; Mou; Gunes e Patras,

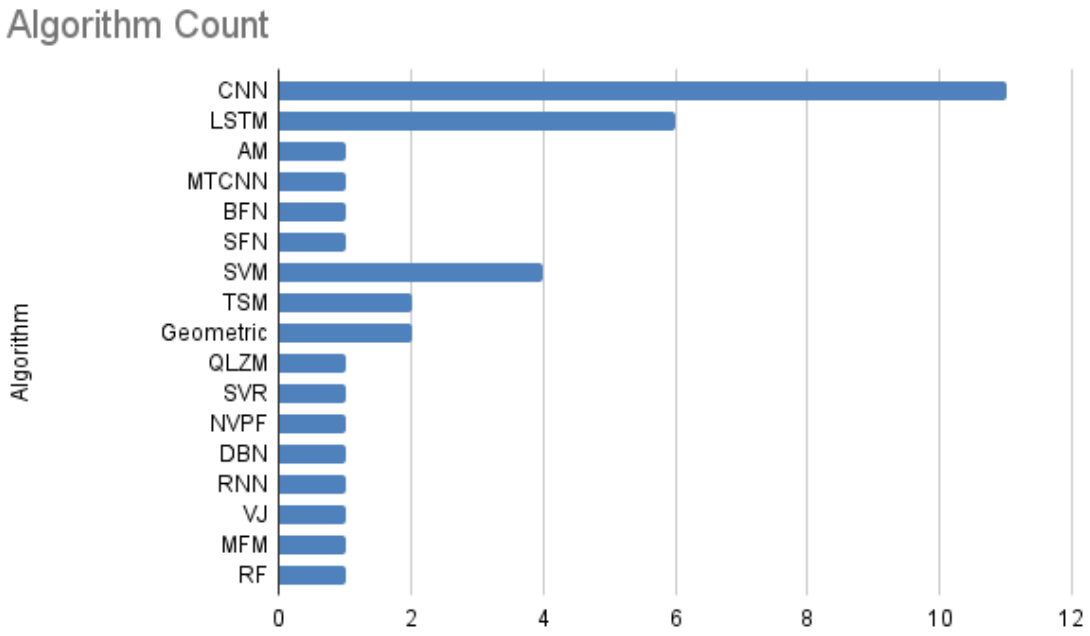
Tabela 4. Quality assessment score designated for each of the articles analyzed, along with the year and country of origin of the paper.

Paper	**Year**	**Country**	**QA**
(Mou; Gunes e Patras, 2019)	2019	England	3.5
(Gupta *et al.*, 2018)	2018	India	3.5
(Wang *et al.*, 2018)	2018	China	3.5
(Yu *et al.*, 2019a)	2019	China	3.5
(Petrova; Vaufreydaz e Dessus, 2020)	2020	France	3.5
(Khan *et al.*, 2018)	2018	USA	3.5
(Guo *et al.*, 2018)	2018	USA	3.5
(Sun *et al.*, 2020)	2020	Netherlands	3.5
(Quach *et al.*, 2022)	2022	Canada	3.5
(Srivastava *et al.*, 2020)	2021	USA	3.5
(Liu *et al.*, 2020)	2020	China	3.0
(Sharma e Mansotra, 2019b)	2019	India	3.0
(Sharma e Mansotra, 2019a)	2019	India	2.5
(Palestra e Pino, 2020)	2020	Italy	2.0

2019) although (Mou; Gunes e Patras, 2019) also employed some other methods such as SVM and LSTM. Figure 2 shows how many papers used the identified algorithms. It's important to point out that in some cases, the authors might have applied multiple algorithms to demonstrate the strengths and weaknesses of each approach, as well as to identify the best fit for a particular problem. For example, they could have used a combination of Geometric Features and Random Forest to compare on a given dataset. In this example, the Geometric Features could be inputted in the Random Forest algorithm to actually make the prediction. In the same way, the combination of multiple datasets could be used to create a hybrid dataset that combines the strengths of each one of them. Finally, authors may have used multiple algorithms to understand the relative importance of different features in a given dataset, as well as to identify potential correlations between features.

From Figure 2 it is evident that CNNs are being largely employed in the context of group-level facial expression recognition, mainly because of their capability to extract features from images automatically and, therefore, classify them in a more assertive way. On the other hand, the results achieved by the CNN algorithms are not necessarily better than some other methods like SVM, QLZM, and even Geometric Features (Mou; Gunes

Figura 2. Count of each algorithm found in the articles analyzed.



e Patras, 2019; Palestra e Pino, 2020).

3.1. CNN

As the name suggests, Convolutional Neural Networks are neural networks that primarily use convolutional layers. These layers take in an input and apply a number of filters to it, producing an output, as known as a feature map (Li *et al.*, 2021b). The filters are generally different for each layer and are learned during training. Initially, a determined number of random filters is applied to the input image and, as the training process evolves, the network can automatically define which are the filters that best describe the input in terms of features relevant to the classification.

In the case of image classification, the input is a 3D tensor, with dimensions for the width, height, and depth (generally the number of color channels). For example, the input for the MNIST dataset would have the dimensions [28, 28, 1]. The output of the layer would have the same width and height as the input, but a different depth. The depth of the output is a function of the number of filters that were applied in the previous layer. In the final layer of the network, the depth of the output is equal to the number of classes that the network is trying to classify (Gu *et al.*, 2018).

The number of filters that will be applied in the first layer of a CNN is usually determined by the problem at hand. The more filters that are used, the more resources (memory, processing power, etc) will be needed. It is common to see CNNs with a number of filters that is a divisor of the depth of the input. For example, if the input has 64 channels, the filters applied in the first layer may be 32, 16, 8, etc.

Besides the convolutional layers, pooling layers are commonly applied in CNNs as well. The purpose of this layer is to reduce the dimensionality of the data, keeping only the most relevant features, therefore optimizing the computational cost to process the network and preventing the network to base its classification on some features that are, in fact, not relevant to the problem. This is done by applying a pooling operator (max pooling, mean pooling, etc) to the output of the convolution layer.

After these two processes: convolution and pooling, normally a conventional Multi-Layer Perceptron (MLP) is used to classify the data. CNNs are widely applied to problems that involve images or videos because of their capability of extracting features from the image and, generally, are very effective at image classification and recognition.

3.2. RNN and LSTM

Recurrent neural networks (RNN) are a type of artificial neural network where connections do not follow a sequential flow, that is, connections between nodes can create a cycle, allowing output from some nodes to affect subsequent input to the same nodes (Yu *et al.*, 2019b). This creates stateful neural networks, which can remember information from previous inputs and use that information to inform future inputs. RNN is a powerful tool for analyzing sequential data such as text, audio, and video.

RNNs are often used for tasks such as language modeling and machine translation, where the order of words is important. Besides that, RNNs can be used for tasks such as sentiment classification and image captioning, where the order of data points is important. There are many different types of RNNs. The most common type is the Long Short-Term Memory (LSTM) network. LSTM is a type of RNN that is capable of learning long-term dependencies. Unlike traditional RNNs, LSTM has a memory cell that can remember information for long periods of time.

The LSTM network is composed of a series of LSTM cells. Each cell has an input gate, output gate, and forget gate. The gates control the flow of information into and out of the cell (Staudemeyer e Morris, 2019). The input gate controls the flow of information from the input to the cell state. The output gate controls the flow of information from the cell state to the output. The forget gate controls the flow of information from the cell state to the forgotten state.

When the LSTM cell is training, the gates are used to control the flow of information into and out of the cell. An optimizer is applied to update the weights of the LSTM cells and, therefore, minimize the loss function.

Usually, LSTM layers are applied along with other layers, such as a fully connected layer, which is used to predict the output, a dropout layer, which is used to prevent overfitting, and a recurrent layer, which is used to learn long-term dependencies (Li *et al.*, 2021a).

LSTMs can be employed in a variety of tasks such as Natural Language Processing (NLP), and machine translation, and, in the field of image processing, LSTMs can be used for image classification, image captioning, and object detection. Image classification can benefit from LSTM networks due to their ability to learn long-term dependencies which can reduce the risk of gradient vanishing that traditional RNN faces.

3.3. Datasets

Supervised learning, which is the case for all the papers taken into account in this study, is highly dependent on the dataset used in the training process. The quality of the dataset used in the training process is determined by the amount of data, the diversity of the data, the number of features, and the quality of the labels. The number of data points and the diversity of data points are essential factors for training a supervised learning algorithm. Furthermore, the quality of the labels and the number of features also play a significant role in the accuracy of the supervised learning algorithm.

The papers all used different datasets for the training process, which were either manually collected or obtained from publicly available datasets. These datasets varied in terms of the number of data points, the diversity of data points, the number of features,

and the quality of labels. The datasets all had different sizes, with some having a few hundred data points and some having thousands of data points. In addition, the number of features and the quality of labels varied among the datasets.

The papers all used different supervised learning algorithms and different evaluation metrics to assess the performance of the algorithms. The algorithms used ranged from traditional machine learning algorithms such as SVM and random forests to deep learning algorithms such as convolutional neural networks and recurrent neural networks. The evaluation metrics used for the algorithms also varied, therefore, it is unfair to compare algorithms based only on the ”best score”provided by the authors. Apart from that, some papers were focused on classifying facial expressions as good, bad or neutral while others based their approach on the 7 base Ekman emotions + the neutral one (Ekman, 1999).

Along the papers, the most cited database was the EmotiW challenge dataset (Srivastava *et al.*, 2020; Sun *et al.*, 2020; Petrova; Vaufreydaz e Dessus, 2020; Wang *et al.*, 2018), which is an audio-video group emotion recognition dataset that contains group videos downloaded from YouTube with creative commons license. This dataset contains 2661 videos for the training set, 756 videos for the validation set and 756 videos for the test set. The data consists of interviews and crowded groups of people talking, for example (Srivastava *et al.*, 2020). It is important to mention that all the labels for this dataset consist of Neutral, Positive and Negative emotions only. On the other hand, it is a dataset of emotions in the wild, that is, faces are sometimes occluded, the environment is different for each sample, and face/body pose may vary according to the video context.

Another implementation of facial expression recognition can be identified in some papers that opted to first extract faces from the scene and, in a second step, recognize the expression individually (Palestra e Pino, 2020; Sharma e Mansotra, 2019a; Guo *et al.*, 2018; Khan *et al.*, 2018). These approaches based their facial expression recognition in image datasets such as CK+ (Lucey *et al.*, 2010a), FER2013 (Goodfellow *et al.*, 2013a), and even from images obtained by the authors from Google Images and Flickr. All the papers, algorithms used, datasets, and the best results achieved by each author are presented in Table 5.

Tabela 5. Each paper analyzed with the respective best score achieved by the authors, along with the algorithms applied and comments about the work.

Paper	**Technology**	**Comments**	**Dataset**	**Data Format**	**Best Score**
(Yu *et al.*, 2019a)	MFM, Deep Bidirectional LSTM	Analyses only three emotions (Negative, Positive and Neutral)	HAPPEI	Image	78.0
(Quach *et al.*, 2022)	CNN, NVPF, TNVPF		Group-level Emotion on Crowded Videos (GECV)	Video	76.12
(Palestra e Pino, 2020)	GF, RF		CK+	Image	94.24
(Sharma e Mansotra, 2019b)	VJ, CNN, RNN, LSTM, SVM		CHEAVD	Video	75.0
(Sharma e Mansotra, 2019a)	SVM, CNN, DBN, LSTM	Analyses facial images and audio.	FER2013	Image	There is no accuracy score in the paper (it appears that the model testing have not been performed yet)
(Wang *et al.*, 2018)	CNN	Analyses only three emotions (Negative, Positive and Neutral)	EmotiW 2018	Video	67.48
(Guo *et al.*, 2018)	CNN, LSTM		FER2013 e GENKI-4K	Image	78.98
(Liu *et al.*, 2020)	CNN, SVM, LSTM, TSM	Analyses facial images and audio.	VGAF + alguns que eles juntaram	Video	76.85
(Gupta *et al.*, 2018)	CNN, AM, MTCNN		Group Affect Database 2.0	Image	64.83
(Khan *et al.*, 2018)	CNN, SFN, BFN		Buscado no Google Images and Flickr	Image	78.39
(Petrova; Vaufreydaz e Dessus, 2020)	CNN	Analyses only three emotions (Negative, Positive and Neutral)	EmotiW 2020 e dataset sintético gerado por eles	Image	59.13
(Sun *et al.*, 2020)	CNN, TSM		EmotiW 2020	Video	71.93
(Srivastava *et al.*, 2020)	CNN		EmotiW 2020	Video	35.0
(Mou; Gunes e Patras, 2019)	GF, QLZM, SVM, SVR, LSTM		AMIGOS (groupDB and individualDB)	Video	94.0

It is noticeable that papers that applied LSTM algorithms were able to achieve some of the best results, although LSTM was always used among other techniques. (Quach *et al.*, 2022; Sharma e Mansotra, 2019b; Sharma e Mansotra, 2019a; Guo *et al.*, 2018; Li e Deng, 2020) and (Mou; Gunes e Patras, 2019) opted to employ LSTM algorithms along with CNNs. This may be explained because of the capability of CNNs to extract features from the images and, in these cases, the LSTM works as a layer in the artificial neural network, therefore, being able to better classify inputs.

The best score achieved by the papers analyzed was 94,24% (Palestra e Pino, 2020). This may be strange when compared with the other papers because of the algorithms applied by the authors. Both the extraction (geometric features) and the classification (random forest) algorithms are not as present in current studies as CNNs or LSTM, for example. Since the classification was able to achieve such a great result, these techniques may be a good choice to be applied in facial expressions and emotions detection problems, although, the robustness of this algorithm implies an increase in training and execution time (Nitze; Schulthess e Asche, 2012).

3.4. Best CNN approaches

Since CNN was the most applied method to solve the problem of this research, it is reasonable to try and deeper understand the applications and results achieved. The two best results achieved by using CNN methods, among the studies presented in this paper, were the ones proposed by (Guo *et al.*, 2018) and (Khan *et al.*, 2018) with 78.98% and 78.39% scores, respectively. Both methods were developed and submitted to the 6th Emotion Recognition in the Wild (EmotiW 2018) Challenge(Dhall *et al.*, 2018).

(Guo *et al.*, 2018) presents a hybrid deep learning network to solve the group-level emotion recognition problem. The classifier proposed by the authors was developed to identify three distinct categories: positive, negative, and neutral emotions. The hybrid approach proposed is a fusion of 8 models in total, among them some well-known models like VGG-FACE, Inception-V2, SE-ResNet-50, and even a LSTM implementation. Some approaches were evaluated separately, but the best result achieved was with the fusion of all models (78.98%). The models were trained and evaluated on a combination of FER-2013(Goodfellow *et al.*, 2013b) and GENKI-4K(Whitehill *et al.*, 2009) datasets.

(Khan *et al.*, 2018) on the other hand, developed an algorithm that works basically in two steps: 1) Face detection and 2) Individual faces predictions. For the first step, the authors applied MTCNN(Zhang *et al.*, 2016) to detect faces from the input images. The MTCNN algorithm is a three-stage cascaded CNN network for joint face detection and landmark localization. For the second step, two different Networks were applied (SFN and BFN), depending on the size of the individual faces detected from step 1. Both SFN and BFN were based on Residual Networks(He *et al.*, 2016). Finally, the group-level prediction was calculated based on the predictions of each face, according to their original size, alleviating the effect of unreliable predictions from smaller background faces. As the previously presented approach, (Khan *et al.*, 2018) designed the algorithm to classify the images into Positive, Negative, and Neutral categories. The dataset used to train and evaluate the model was obtained from Google Images and Flickr based on keyword searches.

3.5. Best LSTM approaches

Other than CNN, LSTM was largely applied in the papers considered in this review. LSTM algorithms are capable of learning long-term dependencies and preventing the vanishing gradient problem (Hochreiter e Schmidhuber, 1997). The two works that

achieved the best results among the ones that applied LSTM algorithms were (Mou; Gunes e Patras, 2019) and (Guo *et al.*, 2018) with 94% and 78.98% respectively.

The approach proposed by (Guo *et al.*, 2018) was already explored in section 3.4. (Mou; Gunes e Patras, 2019), on the other hand, presents an approach to facial detection and emotion recognition that contains LSTM components along with other techniques. Firstly, the authors adopted an SVM-based multi-modal method using dynamic features and conducted experiments on both individual and group videos. The geometric features, used by the algorithm to represent faces, was the facial landmark trajectory and the appearance representations adopted was the extended volume Quantised Local Zernike Moments (QLZM).

The algorithm was trained with both individual and group images with the datasets individualDB and groupDB (Miranda-Correa *et al.*, 2018). In comparison with other algorithms, the authors found that temporal learning models are capable of outperforming non-temporal models in terms of affect recognition, making LSTM a good technique to be applied to this kind of problem. Differently from the other papers explored previously, (Mou; Gunes e Patras, 2019) opted to use subjects' body information as well as face to achieve the best results, mainly to identify the presence of a group of people or a single person in the image (image context).

3.6. Best overall approach

The method proposed by (Palestra e Pino, 2020) is a significant contribution to the field of automatic decoding of facial expressions in video-recorded sessions. The proposed system includes hardware and software components. The main software component, named MCI, was developed using Choreographe (programming tool offered by the robots constructor) and the Python programming language. An humanoid robot and three RGB cameras were responsible for the hardware part of the solution.

The authors opted for a unique approach to implement a system using Geometric Features and a Random Forest classifier, which is different from the majority of the papers explored in the study. This method achieved a high accuracy of 94.24%, which is a remarkable achievement considering that the algorithm was trained and tested using only one dataset, the Extended Cohn-Kanade (CK+) dataset (Lucey *et al.*, 2010b).

The authors believe that the high accuracy rate achieved in their study indicates reliable facial expression recognition, which can provide valuable information to support Human-Robot interactions. The analysis revealed that the system is able to recognize facial expressions from robot-assisted group therapy sessions handling partially occluded faces. The method proposed acts as a mediating tool and appeared to promote the engagement of participants in the training program. The classification results can offer robust information to base its interventions in a group-based therapy session, enhancing the potential for the use of robots in healthcare and rehabilitation settings.

Despite the limitations of the dataset used, the authors have demonstrated the potential of their approach in automatically decoding facial expressions and have provided promising results. Future research could explore the application of the algorithm developed by this study using other datasets to evaluate the generalizability of the proposed method to different populations and broader contexts.

4. Conclusion

Facial expression detection and recognition are not new problems for computer vision and especially for artificial intelligence algorithms, but at the same time, not still a completely solved problem for group-level applications. There has been a good

advance in the last few years, as demonstrated in this review, but algorithms are not quite precise yet. The main challenge is the fact that facial expressions are very subtle and have a lot of variation. Even small changes in the facial expressions can have a big impact on the recognition accuracy. It is also very difficult to train an algorithm that can recognize different expressions in different contexts. On the other hand, scene-based Facial expression recognition in groups is approached not only as a multi-instance approach, but also analysed globally in the image, so that is the reason the number of papers was reduced considerable, taking only the most relevant methods according to the literature review criterion.

Apart from some approaches such as (Mou; Gunes e Patras, 2019) and (Palestra e Pino, 2020), the results still have some improvement to do before becoming great to common life applications. Most of the current approaches are either based on deep learning or on handcrafted features. However, the combination of both has shown promising results. For instance, in (Sharma e Mansotra, 2019b; Sun *et al.*, 2020; Srivastava *et al.*, 2020; Mou; Gunes e Patras, 2019), the authors propose a multimodal approach that combines facial expressions and other inputs such as speech information and body language for better recognition accuracy.

This review provided a general vision of what is being used in terms of group-level facial expression detection and recognition. It was identified that CNNs are largely applied to solve this kind of problem and, besides not having achieved great results yet, with the advance of computing, artificial intelligence, and the creation of new datasets, the performance of these networks is expected to improve in the near future and it is possible that CNNs can benefit from that and show up to be really good classifiers.

It was also possible to identify that CNNs can benefit from LSTM filters to improve the classification results (Mou; Gunes e Patras, 2019; Guo *et al.*, 2018; Yu *et al.*, 2019a). Also, the accuracy level of the algorithms is highly dependent on the dataset on which it was trained and tested on. Therefore, there might be a difference in the accuracy of the algorithms when tested in real-life scenarios.

Finally, although facial expression and emotion detection is not a new topic in computer vision studies, there is still room for improvement, principally for group-level analysis, in both accuracy and complexity levels. With the arise of new deep learning techniques and datasets, it is expected that new advances will be made in the near future. Some of the works analysed in this review, may benefit from larger datasets and greater compute power to elevate the accuracy of their algorithms. Along with these algorithms and methods, it is expected that more research will be carried out to improve the existing results in the field and produce better products for real life scenario applications.

Referências

Ahmad, K. A. B. *et al.* Emerging trends and evolutions for smart city healthcare systems. **Sustainable Cities and Society**, Elsevier, v. 80, p. 103695, 2022.

Ali, G.; Iqbal, M. A.; Choi, T.-S. Boosted nne collections for multicultural facial expression recognition. **Pattern Recognition**, Elsevier, v. 55, p. 14–27, 2016.

Ali, H.; Hariharan, M.; Yaacob, S.; Adom, A. H. Facial emotion recognition using empirical mode decomposition. **Expert Systems with Applications**, Elsevier, v. 42, n. 3, p. 1261–1277, 2015.

Biswas, S.; Sil, J. An efficient expression recognition method using contourlet transform. In: ACM. **Proceedings of the 2nd International Conference on Perception and Machine Intelligence**. [S.l.], 2015. p. 167–174.

Canal, F. Z. *et al.* A survey on facial emotion recognition techniques: A state-of-the-art literature review. **Information Sciences**, Elsevier, v. 582, p. 593–617, 2022.

Dhall, A.; Kaur, A.; Goecke, R.; Gedeon, T. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In: **Proceedings of the 20th ACM International Conference on Multimodal Interaction**. [S.l.: s.n.], 2018. p. 653–656.

Ekman, P. Basic emotions. **Handbook of cognition and emotion**, v. 98, n. 45-60, p. 16, 1999.

Ghasemi, R.; Ahmady, M. Facial expression recognition using facial effective areas and fuzzy logic. In: IEEE. **2014 Iranian Conference on Intelligent Systems (ICIS)**. [S.l.], 2014. p. 1–4.

Goodfellow, I. *et al.* Challenges in representation learning: A report on three machine learning contests. 2013. Disponível em: ⟨http://arxiv.org/abs/1307.0414⟩.

Goodfellow, I. J. *et al.* Challenges in representation learning: A report on three machine learning contests. In: Springer. **International conference on neural information processing**. [S.l.], 2013. p. 117–124.

Graesser, A. C.; Conley, M. W.; Olney, A. Intelligent tutoring systems. **APA educational psychology handbook, Vol 3: Application to learning and teaching.**, American Psychological Association, p. 451–473, 2012.

Gu, J. *et al.* Recent advances in convolutional neural networks. **Pattern recognition**, Elsevier, v. 77, p. 354–377, 2018.

Guo, X.; Zhu, B.; Polanía, L. F.; Boncelet, C.; Barner, K. E. Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions. In: **Proceedings of the 20th ACM International Conference on Multimodal Interaction**. [S.l.: s.n.], 2018. p. 635–639.

Gupta, A.; Agrawal, D.; Chauhan, H.; Dolz, J.; Pedersoli, M. An attention model for group-level emotion recognition. In: **Proceedings of the 20th ACM International Conference on Multimodal Interaction**. [S.l.: s.n.], 2018. p. 611–615.

Happy, S.; Routray, A. Automatic facial expression recognition using features of salient facial patches. **IEEE transactions on Affective Computing**, IEEE, v. 6, n. 1, p. 1–12, 2014.

He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 770–778.

Hochreiter, S.; Schmidhuber, J. Long short-term memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

Huang, X.; Dhall, A.; Goecke, R.; Pietikainen, M. K.; Zhao, G. Analyzing group-level emotion with global alignment kernel based approach. **IEEE Transactions on Affective Computing**, IEEE, 2019.

Hwang, G.-J. A conceptual map model for developing intelligent tutoring systems. **Computers & Education**, Elsevier, v. 40, n. 3, p. 217–235, 2003.

Jain, D. K.; Shamsolmoali, P.; Sehdev, P. Extended deep neural network for facial emotion recognition. **Pattern Recognition Letters**, Elsevier, v. 120, p. 69–74, 2019.

Khan, A. S. *et al.* Group-level emotion recognition using deep models with a four-stream hybrid network. In: **Proceedings of the 20th ACM International Conference on Multimodal Interaction**. [S.l.: s.n.], 2018. p. 623–629.

Kitchenham, B. Procedures for performing systematic reviews. **Keele, UK, Keele University**, v. 33, n. 2004, p. 1–26, 2004.

Li, S.; Deng, W. Deep facial expression recognition: A survey. **IEEE transactions on affective computing**, IEEE, 2020.

Li, W. *et al.* Online capacity estimation of lithium-ion batteries with deep long short-term memory networks. **Journal of power sources**, Elsevier, v. 482, p. 228863, 2021.

Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: analysis, applications, and prospects. **IEEE transactions on neural networks and learning systems**, IEEE, 2021.

Liu, C.; Jiang, W.; Wang, M.; Tang, T. Group level audio-video emotion recognition using hybrid networks. In: **Proceedings of the 2020 International Conference on Multimodal Interaction**. [S.l.: s.n.], 2020. p. 807–812.

Lopes, A. T.; Aguiar, E. D.; Oliveira-Santos, T. A facial expression recognition system using convolutional networks. In: IEEE. **2015 28th SIBGRAPI Conference on Graphics, Patterns and Images**. [S.l.], 2015. p. 273–280.

Lopes, A. T.; Aguiar, E. de; Souza, A. F. D.; Oliveira-Santos, T. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. **Pattern Recognition**, Elsevier, v. 61, p. 610–628, 2017.

Lucey, P. *et al.* The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: IEEE. **2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops**. [S.l.], 2010. p. 94–101.

Lucey, P. *et al.* The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: IEEE. **2010 ieee computer society conference on computer vision and pattern recognition-workshops**. [S.l.], 2010. p. 94–101.

Miranda-Correa, J. A.; Abadi, M. K.; Sebe, N.; Patras, I. Amigos: A dataset for affect, personality and mood research on individuals and groups. **IEEE Transactions on Affective Computing**, IEEE, v. 12, n. 2, p. 479–493, 2018.

Mou, W.; Gunes, H.; Patras, I. Alone versus in-a-group: A multi-modal framework for automatic affect recognition. **ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)**, ACM New York, NY, USA, v. 15, n. 2, p. 1–23, 2019.

Nitze, I.; Schulthess, U.; Asche, H. Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification. **Proceedings of the 4th GEOBIA, Rio de Janeiro, Brazil**, v. 79, p. 3540, 2012.

Palestra, G.; Pino, O. Detecting emotions during a memory training assisted by a social robot for individuals with mild cognitive impairment (mci). **Multimedia Tools and Applications**, Springer, v. 79, n. 47, p. 35829–35844, 2020.

Petrova, A.; Vaufreydaz, D.; Dessus, P. Group-level emotion recognition using a unimodal privacy-safe non-individual approach. In: **Proceedings of the 2020 International Conference on Multimodal Interaction**. [S.l.: s.n.], 2020. p. 813–820.

Quach, K. G. *et al.* Non-volume preserving-based fusion to group-level emotion recognition on crowd videos. **Pattern Recognition**, Elsevier, p. 108646, 2022.

Rebolledo-Mendez, G.; Huerta-Pacheco, N. S.; Baker, R. S.; Boulay, B. du. Meta-affective behaviour within an intelligent tutoring system for mathematics. **International Journal of Artificial Intelligence in Education**, Springer, v. 32, n. 1, p. 174–195, 2022.

Sharma, A.; Mansotra, V. Classroom student emotions classification from facial expressions and speech signals using deep learning. **International Journal of Recent Technology and Engineering**, v. 8, n. 3, p. 6675–6683, 2019. Cited By 0. Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2.0-85073572442&doi=10.35940%2fijrte.C5666.098319&partnerID=40&md5=0e3d9274b894de4866ae80990e9ecc2e⟩.

Sharma, A.; Mansotra, V. Multimodal decision-level group sentiment prediction of students in classrooms. **Int. J. Innov. Technol. Explor. Eng**, v. 8, n. 12, p. 4902–4909, 2019.

Sigov, A.; Ratkin, L.; Ivanov, L. A.; Xu, L. D. Emerging enabling technologies for industry 4.0 and beyond. **Information Systems Frontiers**, Springer, p. 1–11, 2022.

Srivastava, S. *et al.* Recognizing emotion in the wild using multimodal data. In: **Proceedings of the 2020 International Conference on Multimodal Interaction**. [S.l.: s.n.], 2020. p. 849–857.

Staudemeyer, R. C.; Morris, E. R. Understanding lstm–a tutorial into long short-term memory recurrent neural networks. **arXiv preprint arXiv:1909.09586**, 2019.

Sun, M. *et al.* Multi-modal fusion using spatio-temporal and static features for group emotion recognition. In: **Proceedings of the 2020 International Conference on Multimodal Interaction**. [S.l.: s.n.], 2020. p. 835–840.

Văidean, V. L.; Achim, M. V. When more is less: Do information and communication technologies (icts) improve health outcomes? an empirical investigation in a non-linear framework. **Socio-Economic Planning Sciences**, Elsevier, v. 80, p. 101218, 2022.

Wang, K. *et al.* Cascade attention networks for group emotion recognition with face, body and image cues. In: **Proceedings of the 20th ACM international conference on multimodal interaction**. [S.l.: s.n.], 2018. p. 640–645.

Whitehill, J.; Littlewort, G.; Fasel, I.; Bartlett, M.; Movellan, J. Toward practical smile detection. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 31, n. 11, p. 2106–2111, 2009.

Wohlin, C. *et al.* **Experimentation in software engineering**. [S.l.]: Springer Science & Business Media, 2012.

Yu, D.; Xingyu, L.; Shuzhan, D.; Lei, Y. Group emotion recognition based on global and local features. **IEEE Access**, IEEE, v. 7, p. 111617–111624, 2019.

Yu, Y.; Si, X.; Hu, C.; Zhang, J. A review of recurrent neural networks: Lstm cells and network architectures. **Neural computation**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 31, n. 7, p. 1235–1270, 2019.

Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. **IEEE signal processing letters**, IEEE, v. 23, n. 10, p. 1499–1503, 2016.