engineering proceedings

Proceeding Paper

A Survey of Video Analysis Based on Facial Expression Recognition †

Paul Díaz, Elvinn Vásquez * and Pedro Shiguihara *

Department of Information Systems Engineering, Universidad San Ignacio de Loyola, Lima 15024, Peru; paul.diazl@usil.pe

* Correspondence: elvinn.vasquez@usil.pe (E.V.); pshiguihara@usil.edu.pe (P.S.)

† Presented at the II International Congress on the Technology and Innovation in Engineering and Computing, Online, 21–25 November 2022.

Abstract: Verbal language has become the main way to communicate our ideas since the already established linguistic signs allow understanding between people. However, this is not the only way we have to communicate, since nonverbal language, especially facial expressions, usually convey a lot of information. Currently, the use of Convolutional Neural Networks (CNN) has allowed us to identify these emotions more easily through facial expression recognition (FER), which has attracted much attention from various fields of research. In this work, we will provide detailed information about the currently most used dataset and methods for identifying emotions using facial expressions, such as VGG16, ResNet50 and Inception-V3, obtaining a better performance in ResNet-50. This survey shows that the main differences in precision in each architecture are due to the number of images in the datasets used for training and testing.

Keywords: automatic emotion recognition; deep learning; facial expression recognition; convolutional neural network

Citation: Díaz, P.; Vásquez, E.; Shiguihara, P. A Survey of Video Analysis Based on Facial Expression Recognition. *Eng. Proc.* **2023**, *42*, 3. https://doi.org/10.3390/engproc2023042003

Academic Editor: Luis Olivera-Montenegro

Published: 1 August 2023

1. Introduction

Facial expressions are the primary reflection of a person's mental and emotional state. For Albert Mehrabian, considered the father of nonverbal scientific communication, only 7% of real information is transmitted orally, and 38% is conveyed by language auxiliaries, such as the rhythm or speed of speech or tone; on the other hand, 55% of the information that is transmitted is through facial expressions [1].

Ekman and Friesen defined six basic emotions that human beings inherently perceive from our culture, these being sadness, fear, anger, surprise, happiness, and disgust [2]. Although there are currently many studies in different fields such as psychology, psychiatry, health care systems, intelligent systems, human–robot interactions, detection of abnormal driving activities [3], etc., that are related to facial expression recognition (FER) using Convolutional Neural Networks (CNN) [4,5], many of these models are too limited to represent the complexity involved in demonstrating everyday affective activities[5].

This paper presents a survey related to FER that serves as a motivation for researchers and industry players to exploit the full potential of this field, identifying the most current and most used data repositories for this field, as well as the three main steps required in FER systems, focusing primarily on CNN. A comparison of the existing models of Convolutional Neural Networks is also made, aimed at understanding human behavior through the detection of facial expressions.

2. Research Methodology

In order to carry out this study, a search was carried out on the Scopus platform, in which the following query was entered:

TITLE-ABS-KEY ((“CNN” OR “Convolutional neural network” OR “CONVNETS”) AND (“FER” OR “FACIAL EXPRESSION RECOGNITION”) AND (“DATASETS”)) AND (LIMIT-TO (LANGUAGE, “English”)) AND (LIMIT-TO (OA, “all”))

With this, a combined search was performed that involved neural networks, recognition of facial expressions and the term datasets, which is a term of great interest for the present study. From this first query, more terms emerged that were part of the query; this made it possible to increase the amount of information of interest found. The next search was

TITLE-ABS-KEY ((“CNN” OR “Convolutional neural network” OR “CONVNETS”) AND (“DEEP LEARNING”) AND (“FER” OR “FACIAL EXPRESSION RECOGNITION”) AND (“DATASETS”)) AND (LIMIT-TO (OA, “all”)) AND (LIMIT-TO (LANGUAGE, “English”))

After carrying out the queries shown, 84 documents were shown, of which 25 were selected that will contribute greatly to what we want to make known in this survey.

3. Database

Although there are several datasets that can be used, when designing an optimal FER system it is important to have a large dataset that has a large amount of trained and labeled data, including environmental and population variations. In this section, we will address the existing and most used databases in the different articles reviewed, as shown in Table 1.

- (CK+) [6]: This is the most used laboratory dataset when it comes to emotion detection; it has a total of 593 video sequences, which have a duration of 10 min at 60 fps.
- FER 2013 [7]: This is a dataset compiled by the Google image search API; due to this, it is a large-scale dataset without restrictions, containing 28,709 training images, 3589 validation images and 3589 test images. This is the second most used dataset after CK+.
- The Japanese Female Facial Expressio (JAFFE) [4]: This is a laboratory-controlled image database, containing 213 expression samples; in this, 10 women have 34 images for each of the six basic facial expressions and 1 image with a neutral expression.

Table 1. List of most commonly used datasets for facial expression recognition, each with its access link, references and a detailed description of the images and videos contained in the dataset.

DataSet	Reference	Description
CASME II http://casme.psych.ac.cn/casme/e2 (Accessed on 15 March 2023)	Singh (2019) [5]	This dataset contains a total of 3000 facial movements and 247 microexpressions, which are divided into positive, negative and neutral categories: anger: 37 negative videos; contempt: 17 negative videos; disgust: 31 negative videos; fear: 12 negative videos; sadness: 23 negative videos; joy: 26 positive videos; surprise: 25 positive videos; and neutral with a total of 46 negative videos.
SMIC	Singh (2019) [5]	The SMIC dataset contains 164 spontaneous microexpressions from 16 participants. Joy: 69 videos; sadness: 75 videos; anger: 69 videos; disgust: 72 videos; surprise: 72 videos; and fear: 72 videos.
KDEF	Akhand (2021) [4], Shehu (2022) [8]	The dataset contains a total of 4900 static images divided into 6 classes of expressions. Anger: 50 images; fear: 40 images; disgust: 40 images; happiness: 69 images; sadness: 59 images; and surprise: 28 images.
CK+https://www.kaggle.com/datasets/davilsena/ckdataset (Accessed on 20 February 2023)	Lucey (2010) [6], Li (2020) [9], Cai (2018) [10], Shehu (2022) [8]	This dataset contains a total of 920 grayscale images, each with a resolution of 2304 pixels (48 × 48). The dataset is split into 80% for training, 10% for public testing, and 10% for private testing. The number of images for each expression is as follows: anger: 45 images; disgust: 59 images; fear: 25 images; happiness: 69 images; sadness: 28 images; surprise: 83 images; neutral: 593 images; and contempt: 18 images.

Table 1. *Cont.*

DataSet	Reference	Description
JAFFE https://zenodo.org/record/3451524#.Y4fh2KLMKbU (Accessed on 10 March 2023)	Akhand (2021) [4], Penny, S. (1998) [11], Li (2020) [9]	This dataset contains a total of 213 grayscale images of Japanese female facial expressions, with a resolution of 256 × 256 pixels. The number of images for each expression is as follows: anger: 30 images; happiness: 31 images; sadness: 31 images; surprise: 31 images; fear: 30 images; disgust: 31 images; and neutral: 29 images.
FER2013	Monica, B. (2013) [7], Melinte, D.O. (2020) [12], Li (2020) [9]	This dataset contains a total of 35,887 grayscale images of human faces labeled with 6 emotions; each image has a resolution of 48 × 48 pixels. The number of images for each emotion is as follows: anger: 3180 images; fear: 1005 images; happiness: 7264 images; sadness: 5178 images; surprise: 2114 images; and neutral: 17,146 images
AffectNet	Mollahosseini (2019) [13]	This dataset consists of over one million images in JPEG format, where each image is labeled with one or multiple facial expressions. The category of neutral has 459,652 images, happiness 224,500 images, sadness 1,113,997 images, anger 44,418 images, fear 33,345 images, surprise 28,881 images, disgust 3102 images, and contemplation 1211 images
MMI https://mmifacedb.eu/ (Accessed on 12 March 2023)	Pantic (2005) [14], Cai (2018) [10]	This dataset consists of over 2900 high-resolution videos, in which the identified emotions are joy, sadness, anger, surprise, fear, and disgust.
AFEW https://ibug.doc.ic.ac.uk/resources/afew-va-database/ (Accessed on 24 February 2023)	Dhall (2012) [15]	This dataset contains approximately 1809 videos of facial expressions, including videos of individuals expressing six basic emotions: happiness (571 videos), sadness (527 videos), anger (485 videos), surprise (465 videos), fear (215 videos), and disgust (256 videos).
Yale B https://www.kaggle.com/datasets/olgabelitskaya/yale-face-database (Accessed on 7 March 2023)	Bendjillali (2022) [16]	This dataset contains a total of 165 GIF images from 15 subjects, with 11 images per subject. The facial expressions used in the images include happiness, neutral, sadness, drowsy, and surprised
CMU PIE https://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html(Accessed on 1 March 2023)	Bendjillali (2022) [16]	This dataset contains over 750,000 images of 337 individuals. The subjects were photographed from 15 viewpoints and under 19 lighting conditions while displaying a variety of facial expressions. Additionally, high-resolution frontal images were also acquired. In total, the database contains over 305 GB of facial data, including 40 images of happiness expressions, 38 of sadness, 46 of anger, 51 of surprise, 42 of disgust, 38 of fear, 20 of neutral expressions, and 48 of smiles

4. Facial Expressions Recognition

The FER system includes the main stages of facial image preprocessing, feature extraction, and classification. Each one has different techniques; however, for this study only methods belonging to CNN in the second era of FER will be developed.

4.1. Architectures

The three architectures chosen for this comparison are VGG, ResNet, and Inception. According to [17], the first one stood out in 2014 because it used small-sized kernels and had a homogeneous topology, placing it in the "Spatial Exploitation" category. On the other hand, ResNet stood out for its identified mapping-based skip connections with residual learning in 2016, entering the "Depth + Multi-Path" category. And as for Inception, in 2015 it had its main role because of its identified mapping-based skip connections with residual learning, which made it belong to the "Depth + Width" category.

4.2. Comparative

After analyzing the bibliography, a precision comparison of three of the most used architectures has been made, i.e., VGG-16, ResNet-50, and Inception-V3 [16,18]. All these architectures were evaluated with the training and testing of various datasets that allow us

to obtain a more precise vision of the differences between them. Given this, it is evident that the VGG-16 model has better precision when trained and tested with the CMU PIE dataset, since it obtained 97.41%. Likewise, the ResNet-50 model also obtained a higher precision in CMU PIE of 99.53%. Finally, regarding the Inception-v3 model, it has a 99.89% precision in the same dataset as the previous ones.

From Table 2, it is determined that the most stable architecture is ResNet, since the results obtained with the various datasets have a determination coefficient of 0.82, which indicates that the values are close to the regression line, while VGG has a coefficient of 0.69.

Table 2. Technical specifications of the most commonly used datasets for facial expression recognition, including the number of classes (expressions or microexpressions), image and video resolutions (if the dataset contains them), video frames per second (fps), video duration, and whether the dataset is public.

Dataset	# Classes	Img. Res	Vid. Res	Fps	Duration	Public?
CASME II	8	640 × 480 px–1280 × 960 px	640 × 480 px–1280 × 960 px	60	5 s	Yes
SMIC	6	320 × 240 px–640 × 480 px	640 × 480 px–1280 × 960 px	25–30	2 h 30 min	No
KDEF	6	490 × 640 px	-	-	-	Yes
CK+	8	48 × 48 px		-	-	Yes
JAFFE	7	256 × 256 px	-	-	-	Yes
FER2013	6	48 × 48 px	-	-	-	Yes
AffectNet	8	224 × 224 px–512 × 512 px		-	-	Yes
MMI	6	640 × 480 px–800 × 600 px–1280 × 960 px	640 × 480 px–800 × 600 px–1280 × 960 px	30	-	Yes
AFEW	6	640 × 480 px–1920 × 1080 px	640 × 480 px–1920 × 1080 px	25	1 s to 6 s	Yes
Yale B	5	192 × 168 px	-	-	-	Yes
CMU PIE	8	320 × 240 px–640 × 480 px	-	-	-	Yes

On the other hand, in Table 3 comparing the results by dataset, it is found that the one with the best results is the CMU PIE [16], while the one that obtained the lowest results was KDEF [8]. This is because the first dataset contains 41,368 images of faces, while the second has only 4900 images. This shows that the greater the number of images in the datasets, the greater the precision of the trained architectures.

Table 3. The recognition results of VGG-16, ResNet-50, and Inception-V3 obtained from testing [18].

Dataset	VGG-16	ResNet-50	Inception-v3	Cite
Extended Yale B	97.28	98.35	99.44	[16]
CMU PIE	97.41	99.53	99.89	[16]
CK+	88.16	97.43		[8]
KDEF	63.57	86.05		[8]
EAFE	75.45	80.53	78.38	[18]

5. Conclusions

From this work, we can conclude that out of a total of 25 articles reviewed, 13 were useful for identifying the most used datasets in FER systems; CK+ is used by 31%, followed by JAFFE and FER2013 with 23%, while the rest of the datasets are used in between 15% and 8%.

In addition, the three most used convnet models were analyzed, from which it was found that the dataset with which they are trained and tested can greatly influence the accuracy of the model. A clear example is the VGG-16 architecture, which obtained a precision of 97.41% with the CMU PIE dataset, while when trained with the KDEF dataset it obtained a precision of 63.57%, representing a 33.84% difference. This difference is mainly due to the number of images used in the datasets with which they were trained and tested.

Author Contributions: Conceptualization, E.V. and P.D.; methodology, E.V. and P.D.; validation, E.V., P.D., and P.S.; investigation, E.V. and P.D.; resources, E.V.; writing—review and editing, P.S.; visualization, P.D.; supervision, P.S.; project administration, P.S.; funding acquisition, P.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not aplicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Roopa, S.N. Research on face expression recognition. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *8*, 88–91. [CrossRef]
2. Ekman, P.; Friesen, W.V. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*; Prentice-Hall: Cambridge, MA, USA, 2003.
3. Shi, M.; Xu, L.; Chen, X. A Novel Facial Expression Intelligent Recognition Method Using Improved Convolutional Neural Network. *IEEE Access* **2020**, *8*, 57606–57614. [CrossRef]
4. Akhand, M.A.H.; Roy, S.; Siddique, N.; Kamal, M.A.S.; Shimamura, T. Facial Emotion Recognition Using Transfer Learning in the Deep CNN. *Electronics* **2021**, *10*, 1036. [CrossRef]
5. Singh, S.; Nasoz, F. Facial Expression Recognition with Convolutional Neural Networks. In Proceedings of the 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 6–8 January 2020; pp. 324–328. [CrossRef]
6. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101. [CrossRef]
7. Monica, B.; Marco, M.; Lakhmi, C.J. *Neural Information Processing* ; Springer: Berlin/Heidelberg, Germany, 2013.
8. Shehu, H.A.; Browne, W.N.; Eisenbarth, H. An anti-attack method for emotion categorization from images [Formula presented]. *Appl. Soft Comput.* **2022**, *128*, 109456. [CrossRef]
9. Li, J.; Jin, K.; Zhou, D.; Kubota, N.; Ju, Z. Attention mechanism-based CNN for facial expression recognition. *Neurocomputing* **2020**, *411*, 340–350. [CrossRef]
10. Cai, J.; Meng, Z.; Khan, A.S.; Li, Z.; O'Reilly, J.; Tong, Y. Island Loss for Learning Discriminative Features in Facial Expression Recognition. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face I& Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018. [CrossRef]
11. Penny Storms. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; IEEE Computer Society: Los Alamitos, CA, USA, 1998.
12. Melinte, D.O.; Vladareanu, L. Facial expressions recognition for human–robot interaction using deep convolutional neural networks with rectified adam optimizer. *Sensors* **2020**, *20*, 2393. [CrossRef] [PubMed]
13. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Trans. Affect. Comput.* **2019**, *10*, 18–31. [CrossRef]
14. Pantic, M.; Valstar, M.; Rademaker, R.; Maat, L. Web-based database for facial expression analysis. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6 July 2005; Volume 2005, pp. 317–321. [CrossRef]
15. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimed.* **2012**, *19*, 34–41. [CrossRef]
16. Bendjillali, R.I.; Beladgham, M.; Merit, K.; Taleb-Ahmed, A. Illumination-robust face recognition based on deep convolutional neural networks architectures. *Indones. J. Electr. Eng. Comput. Sci.* **2020**, *18*, 1015–1027. [CrossRef]
17. Li, S.; Guo, L.; Liu, J. Towards East Asian Facial Expression Recognition in the Real World: A New Database and Deep Recognition Baseline. *Sensors* **2022**, *22*, 8089. [CrossRef]
18. Bhatt, D.; Patel, C.; Talsania, H.; Patel, J.; Vaghela, R.; Pandya, S.; Modi, K.; Ghayvat, H. CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope. *Electronics* **2021**, *10*, 2470. [CrossRef]