applied sciences

MDPI

Article

Emotion Identification in Movies through Facial Expression Recognition

João Almeida [1,2], Luís Vilaça [1,3], Inês N. Teixeira [1,2] and Paula Viana [1,3,*]

[1] INESC TEC, 4200-465 Porto, Portugal; j.almeida@fe.up.pt (J.A.); luis.m.salgado@inesctec.pt or 1121405@isep.ipp.pt (L.V.); ines.f.teixeira@inesctec.pt or up201104124@edu.fe.up.pt (I.N.T.)
[2] Faculty of Engineering, University of Porto, 4200-465 Porto, Portugal
[3] School of Engineering, Polytechnic of Porto, 4200-072 Porto, Portugal
[*] Correspondence: paula.viana@inesctec.pt or pmv@isep.ipp.pt

Abstract: Understanding how acting bridges the emotional bond between spectators and films is essential to depict how humans interact with this rapidly growing digital medium. In recent decades, the research community made promising progress in developing facial expression recognition (FER) methods. However, no emphasis has been put in cinematographic content, which is complex by nature due to the visual techniques used to convey the desired emotions. Our work represents a step towards emotion identification in cinema through facial expressions' analysis. We presented a comprehensive overview of the most relevant datasets used for FER, highlighting problems caused by their heterogeneity and to the inexistence of a universal model of emotions. Built upon this understanding, we evaluated these datasets with a standard image classification models to analyze the feasibility of using facial expressions to determine the emotional charge of a film. To cope with the problem of lack of datasets for the scope under analysis, we demonstrated the feasibility of using a generic dataset for the training process and propose a new way to look at emotions by creating clusters of emotions based on the evidence obtained in the experiments.

Keywords: facial expression recognition; emotion analysis; machine learning; deep learning; computer vision

check for updates

Citation: Almeida, J.; Vilaça, L.; Teixeira, I.N.; Viana, P. Emotion Identification in Movies through Facial Expression Recognition. *Appl. Sci.* **2021**, *11*, 6827. https://doi.org/10.3390/app11156827

Academic Editor: Theodoros Giannakopoulos

Received: 18 June 2021
Accepted: 21 July 2021
Published: 25 July 2021

1. Introduction

Films are rich means of communication produced for cultural and entertainment purposes. Audio, text, and image work together to tell a story, trying to transmit emotional experiences to the audience. The emotion dimension in movies is influenced by the filmmakers' decisions in film production, but it is especially through acting that emotions are directly transmitted to the viewer. Characters transmit their emotions through the actors' facial expressions, and the audience experiences an emotional response.

Understanding how this bond between represented emotion and perceived emotion is created can give us concrete information on human interaction with this rapidly growing digital medium. This can integrate large film streaming platforms and be used for information retrieval concerning viewer experience, quality review, and for the improvement of state-of-the-art recommendation systems. Additionally, this matter falls into the field of affective computing, which is an interdisciplinary field that studies and develops systems that can recognize, interpret, process and simulate human affection. Therefore, emotional film perception could also be a contributing factor for creating affective movie streaming platforms.

Specifically, the challenge lies in answering the following question: "What emotion does this particular content convey?" This is studied in detail in the subfield of emotion and sentiment analysis by analyzing different modalities of content. More specifically, text-based sentiment analysis has been the reference in this area, with the use of natural language processing (NLP) and text analysis techniques for the extraction of the sentiment that a text conveys. A common application of these techniques is social network text analysis and e-commerce online reviews analysis, due to the proven added value to companies and organizations. Advances in computer vision (CV) and machine learning (ML) have, however, shifted the focus of this field by starting to leverage visual and aural content instead of only considering unimodal text-based approaches. The advantage of analyzing the three media present in movies by only assessing text is the possibility of taking into account the character behavior context: it is possible to combine visual and sound cues to better identify the true affective state represented in a film.

When analyzing movies, other stylistic characteristics can be used to improve the accuracy of emotion recognition. For instance, it is common practice to use camera close-ups to evoke intense emotions in the audience. Although the research community has made promising progress in developing facial expression recognition methods, the application of current approaches on the complex nature of a film, where there is a strong variation in lighting and pose, is a problem far from being solved.

This work aimed to investigate the applicability of current automatic emotion identification solutions in the movie domain. We intended to gather a solid understanding of how emotions are addressed in the social and human sciences and discuss how emotional theories are adapted by classification models with deep learning (DL) and machine learning (ML). Taking into account the relevant available datasets, we selected two datasets for our experiments: one containing both posed (i.e., in controlled environments) and spontaneous (i.e., unplanned settings) image web samples, and another that contains images sampled from movies (i.e., with posed and spontaneous expressions). We benchmarked existing CNN architectures with both datasets, initializing them with pre-trained weights from ImageNet. Due to the inclusion of images in uncontrolled environments, the obtained results fall below what would be expected for this task. Hence, we discuss the reliability of multi-class classification models, their limitations, and possible adjustments to achieve improved outcomes. Based on the findings obtained in other multi-media domains that also explore affective analysis, we propose to reduce the number of discrete emotions based on the observation that overlap between classes exists and that clusters can be identified.

What remains of this article is structured as follows: Section 2 defines the problem this work intended to tackle and presents the related work regarding it; Section 3 provides a synthesis of the conducted study, including a detailed definition of the evaluation and analysis methodology with a description of the methods and datasets used to address it; Section 4 depicts and discusses the obtained results; Section 5 concludes by pointing out future paths to be pursued for automatic emotion identification.

2. Related Work

2.1. Emotion Description and Representation

In their exploratory work, Paul Ekman argued [1] that facial expressions are universal and provide sufficient information to predict emotions. His studies suggest that emotions evolved through natural selection into a limited and discrete set of basic emotions: anger, disgust, fear, happiness, sadness, and surprise. Each emotion is independent of the others in its behavioral, psychological and physiological manifestations, and each is born from the activation of unique areas in the central nervous system. The criterion used was the assumption that each primary emotion has a distinct facial expression that is recognized even between different cultures [2]. This proposal set the grounds to other studies that tried to expand the set of emotions to non-basic ones, such as fatigue, anxiety, satisfaction, confusion, or frustration (e.g., Ortony, Clore and Collins's Model of Emotion (OCC)) [3–5].

To bridge emotion theory and visual observations from facial expressions, Ekman also proposed a Facial Action Coding System (FACS) [6]. FACS is an anatomically based system used to describe all visually discernible movement of face muscles, from which it is possible to objectively measure the frequency and intensity of facial expressions using a scale based on Action Unit (AU), i.e., the smallest distinguishable unit of measurable facial movement, such as brow lowering, eyes blinking or jaw dropping. The system has a total of 46 action units, each with a five-point ordinal scale, used to measure the degree of contraction. FACS is strictly descriptive and does not include an emotion correspondence. Therefore, the same authors proposed an Emotional Facial Action Coding System (EMFACS) [7] based on the six-basic discrete emotion model, thus making a connection between emotions and facial expressions. Recent studies have proposed a new classification system based on simple and compound emotions [8].

Despite being the dominant theory in psychology and neuroscience research, recent studies have pointed out some limitations in the six-basic emotion's model. Certain facial expressions are associated with more than one emotion, which suggests that the initial proposed taxonomy is not adequate [9]. Other studies suggest that there is no correlation between the basic emotions and the automatic activation of facial muscles [10], while other claims suggest that this model is culture-specific and not universal [11]. These drawbacks caused the emergence of additional methods that intend to be more exhaustive and universally accepted regarding emotion classification.

Some studies have assessed people's difficulty in evaluating and describing their own emotions, which points out that emotions are not discrete and isolated entities, but rather ambiguous and overlapping experiences [12]. This line of thought reinforced a dimensional model of emotions, which describes them as a continuum of highly interrelated and often ambiguous states. The model that gathered the most consensus among researchers—the Circumplex Model of Emotion—argues that there are two fundamental dimensions: valence, which represents the hedonic aspect of emotion (that is, how pleasurable it is for the human being), and arousal, which represents an enthusiasm or tension dimension, (i.e., the energy level) [13]. Hence, each emotion is represented using coordinates in a multi-dimensional state.

Other approaches propose either bi-dimensional [14] or tri-dimensional (e.g., Pleasure-Arousal (PA) and Pleasure-Arousal-Dominance (PAD) [15]) models for representing emotions. The utility of a third dimension remains unclear, as several studies revealed that the valence and arousal axes are sufficient to model emotions, particularly when handling emotions induced by videos [16]. However, Fontaine, after proposing a model with four dimensions, concluded that the optimal number of dimensions depends on the specificity of the targeted application/study [17].

The advantages of a dimensional model compared with a discrete model are the accuracy in describing emotions, by not being limited to a closed set of classes, and a better description of emotion variations over time, since they are not realistically discrete, but rather continuous.

Motivated by the dispersion of classification methods across emotional datasets, some studies have investigated the potential mapping between discrete/categorical and dimensional theories. In 2011, a first linear mapping between PAD and OCC emotion models [18] was proposed [19]. Nevertheless, it was based on theoretical assumptions, instead of using evidence-based studies. In 2018, a new study elaborated a mapping between Ekman's six basic emotions and the PAD model [20] by cross-referencing information of lexicons (i.e., Affective Norms for English Words (ANEW) [21] and Synesketch [22] lexicons) annotated in both models. Furthermore, they also derived a PA mapping using Nencki Affective Word List (NAWL) [23,24].

Using these lexicons datasets (ANEW, NAWL), an exploratory data analysis indicated the apparent formation of emotion clusters in the PA model: emotions with negative connotation have high overlap, specially between anger and sadness, while neutral and happy form individual clusters in the high-valence and medium-arousal, and in the low-arousal and low-valence regions, respectively. A similar analysis, in the aural domain [25], concluded that similar cluster regions exist, particularly in the happiness emotion and the overlap of "negative" emotions.

2.2. Facial Expression Recognition

Facial expression recognition (FER) systems use biometric markers to detect emotion in human faces. Since 2013, international competitions (such as FER2013 [26] and EmotiW [27]) have changed the facial expressions recognition paradigm by providing a significant increase in training data. These competitions introduced more unconstrained datasets which led to the transition of the study from controlled environments in the laboratory to more unrestrained settings.

2.2.1. Datasets for FER

Datasets are the fundamental piece in any machine learning application and several relevant datasets have been made available and used in most of the FER experiments.

1. Acted Facial Expressions In The Wild (AFEW) [28]: Dataset consists of 1809 video segments extracted from movies. It is labeled with Ekman's discrete emotional model plus a neutral emotion class. The labeling process uses a recommendation system to suggest video clips to a human labeler through their subtitles. The annotations contain the perceived emotions and information regarding the actors present in the clip, such as their name, head-pose and age.
2. AFEW-VA [29]: AFEW-VA is an extension of the AFEW dataset, from which 600 videos were selected and annotated for every frame using the dimensional emotion model (valence and arousal) for every facial region, which is described using 68 facial landmarks.
3. AffectNet [30]: AffectNet contains more than 1 million facial images collected from the web by making queries using 1250 keywords related to emotions in six different languages. The entire database was annotated in the dimensional model (valence and arousal), and half of the database was manually annotated in both the categorical (with the eight labels: neutral, happy, sad, surprise, fear, disgust, anger, contempt, none, uncertain and non-face) and dimensional models.
4. Aff-Wild2 [31]: The extended Aff-Wild database contains 558 videos annotated in continuous emotions (dimensional model—valence and arousal), using different AUs, and a set of 18 discrete FER classes, which also contain the six basic emotions.
5. AM-FED+ [32]: The Extended Dataset of Naturalistic and Spontaneous Facial Expressions Collected in Everyday Settings (AM-FED+) consists of 1044 facial videos recorded in real-world conditions. All the videos have automatically detected facial landmark locations for every frame and 545 of the videos were manually FACS coded. A self-report of "liking" and "familiarity" responses from the viewers is also provided.
6. CK+ [33]: The Extended Cohn-Kanade (CK+) is the most widely adopted laboratory-controlled dataset. The database is composed of 593 FACS coded videos, 327 of which are labeled with the six basic expression labels (anger, disgust, fear, happiness, sadness and surprise) and contempt. CK+ does not provide specific training, validation and test sets.
7. EmotioNet [8]: The EmotioNet database includes 950,000 images collected from the Web, annotated with AU, AU intensity, basic and compound emotion category, and WordNet concept. The emotion category is a set of classes extended from the discrete emotion model. Emotion categories and AUs were annotated using the algorithm described in [8].
8. FER2013 [26]: FER2013 was introduced in the ICML 2013 Challenges in Representation Learning, and consists of 48 × 48 pixel grayscale images of faces. The images were collected using Google's image search Application Programming Interface (API), in which the facial region is centered, resized and cropped to roughly occupy the same amount of space in each image. The database is composed of 28,709 training, 3589 validation and 3589 test images with seven emotion labels: anger, disgust, fear, happiness, sadness, surprise and neutral.
9. JAFFE [34]: The Japanese Female Facial Expression (JAFFE) is one of the first facial expression datasets. It contains seven facial expressions (i.e., the labels from the dis-

crete emotion model and a neutral label). The database is composed of 253 grayscale images with a resolution of 256 × 256 px.

10. KDEF [35]: The Karolinska Directed Emotional Faces (KDEF) is a set of 4900 pictures annotated using a model with six facial expression classes (happy, angry, afraid, disgusted, sad, surprised and neutral). The set of pictures registers 70 subjects (35 men and 35 women), viewed from five different angles.
11. MMI [36,37]: MMI Facial Expression is a laboratory-controlled dataset and has over 2900 videos of 75 subjects. Each video was annotated for the presence of AUs and the six basic expressions plus neutral. It contains recordings of the full temporal pattern of a facial expression, from the neutral state to the peak expression, and back to neutral.
12. OULU-CASIA [38]: Contains 2880 videos categorized into six basic expressions: happiness, sadness, surprise, anger, fear, disgust. The videos were recorded in a laboratory environment, using two different cameras (near-infrared and visible light) under three different illumination conditions (normal, weak and dark conditions). The first eight frames of each video correspond to the neutral class, while the last frame contains the peak expression.
13. RAF-DB [39,40]: Real-world Affective Faces Database (RAF-DB) contains 29,672 facial images downloaded from the Internet. The dataset has a crowdsourcing-based annotation with the six basic emotions, a neutral label, and twelve compound emotions. For each image, facial landmarks, bounding box, race, age range and gender attributes are also available.
14. SFEW [41]: Static Facial Expressions in the Wild (SFEW) contains frames selected from AFEW. The dataset was labeled using the discrete emotion model plus the neutral class. It contains 958 training, 372 testing and 436 validation samples. The authors also made available a pre-processed version of the dataset with the faces aligned in the image. SFEW was built following a Strictly Person Independent (SPI) protocol, therefore the train and test datasets contain different subjects.

Table 1 provides an overview of the FER databases. At the moment, to the best of our knowledge, AFEW [28] (and its extensions to SFEW [41] and AFEW-VA [29]) is the only facial expression dataset in the movie domain, which poses a considerable obstacle for data-based methods given its very limited size. An alternative would be joining datasets from other domains, but there is some evidence that increasing the size of databases in training resulted in small increases in cross-domain performance [42]. Additionally, there is a huge variability of annotations between datasets, which complicates generalization across domains.

FER datasets share several properties, namely the shooting environment and the elicitation method. The shooting environment is closely related to the data quality and thus to the performance of deep FER systems. Laboratory-controlled shooting environments provide high-quality image data where illumination, background and head poses are strictly imposed. However, building these datasets is a time-consuming process and consequently, they are limited in the number of samples. In-the-wild settings, on the other hand, are easier to collect but prove to be challenging when attempting to achieve high-performance deep learning models.

The elicitation method refers to the way that the person pictured in an image portrayed the supposed emotion. Posed expression datasets, in which facial behavior is deliberately performed, are often exaggerated, increasing the differences between classes and making the images easier to classify. Spontaneous expression datasets are collected under the guarantee of containing natural responses to emotion inductions, better reflecting a real-word scenario. Datasets that were collected from the Web or movies normally include both posed and spontaneous facial behavior. Additionally, the discrete model of emotions predominates FER datasets.

Table 1. Principal databases used in FER systems (C. = color; Res. = resolution; G = grayscale; RGB = RGB-colored; P = posed (expression); S = spontaneous (expression); BE = basic emotion; AU = action unit; CE = compound emotion; FL = facial landmark).

Database	Sample (C.Res.)	Subjects	Source	Annotation
AFEW [41]	1809 videos (RGB N/A)	330	Movie ∣ P & S	6 BEs + Neutral
AFEW-VA [29]	600 videos (RGB N/A)	240	Movie ∣ P & S	Valence, Arousal, FLs
AffectNet [30]	450,000 images (RGB 425 × 25)	N/A	Web ∣ P & S	6 BEs + Neutral
Aff-Wild2 [31]	558 videos (RGB 1454 × 890)	458	Web ∣ S	Valence, Arousal
AM-FED+ [32]	1044 videos (RGB 320 × 240)	1044	Web ∣ S	11 AUs, FLs, Liking
CK+ [33]	593 videos (RGB 640 × 480)	123	Lab ∣ P	7 BEs + contempt, AUs, FLs
EmotioNet [8]	950,000 images (RGB N/A)	N/A	Web ∣ P & S	12 AUs, 23 BE and CEs
FER2013 [26]	35,887 images (G 48 × 48)	N/A	Web ∣ P & S	6 BEs + Neutral
KDEF [35]	4900 images (RGB 562 × 762)	70	Lab ∣ P	6 BEs + Neutral
JAFFE [34]	213 images (G 256 × 256)	10	Lab ∣ P	6 BEs + Neutral
MMI [36,37]	2900 videos (RGB 720 × 576)	75	Lab ∣ P	6 BEs + Neutral, AUs
OULU-CASIA [38]	2880 videos (RGB 320 × 240)	80	Lab ∣ P	6 BEs
RAF-DB [39,40]	26,672 images (RGB N/A)	N/A	Web ∣ P & S	6 BEs + Neutral, 42 FLs and 12 CEs
SFEW [41]	1766 images (RGB N/A)	95	Movie ∣ P & S	6 BEs + Neutral

2.2.2. FER Methodologies

Table 2 demonstrates State of the Art (SoA) approaches and results on the most widely evaluated categorical datasets. SoA approaches achieve over 90% of accuracy in CK+ [33] and JAFFE [34], which is justified since they are datasets with laboratory-controlled ideal conditions. However, datasets with subjects who perform spontaneous expressions under "in-the-wild" scenario conditions, such as FER2013 [26] and SFEW [41], have less satisfactory results.

Table 2. FER approaches and results on widely evaluated datasets.

Study (Year)	Approach	Acc (%) (No. Classes)
	CK+ [33]	
[43] (2019)	FAN	99.69 (7)
[44] (2016)	CNN	98.9 (6)
[45] (2017)	CNN	98.62 (6)

Table 2. *Cont.*

Study (Year)	Approach	Acc (%) (No. Classes)
	FER2013 [26]	
[46] (2016)	CNN-VGG	72.7 (7)
[45] (2017)	CNN	72.1 (7)
[46] (2016)	CNN-Inception	71.6 (7)
	JAFFE [34]	
[47] (2019)	SVM	97.10 (7)
[48] (2019)	2channel-CNN	95.8 (7)
[49] (2019)	ATN	92.8 (7)
	SFEW [41]	
[50] (2015)	CNN	55.96 (7)
[51] (2015)	CNN	53.9 (7)
[52] (2019)	CNN (ACNN)	51.72 (7)
	OULU-CASIA [38]	
[53] (2018)	GAN + CNN	88.92 (6)
[52] (2019)	CNN (ACNN)	58.18 (6)

As shown in Table 2, CNN-based approaches are the foundation of SoA results and can be applied to FER tasks to achieve consistent performances. These SoA methods/models are derived from traditional DL architectures, which use well-known backbones for feature extraction (e.g., VGG, ResNet).

Directly using these standard feature extractors and fine tuning the *softmax* layer can contribute to softening the FER's small dataset problem. However, it creates a bottleneck because it relies on a predefined feature space. This issue is commonly tackled by using multistage fine-tuning strategies based on different combinations of the training dataset to enhance their performance [54] or by using facial recognition feature extractors and regularizing them with facial expression information [55].

To increase the power of representations for FER, several works have proposed novel architectures for increasing the depth of multi-scale features [56] or increasing the level of supervision in embedded representations [57]. Additionally, common limitations associated with *softmax* are caused by inter-class similarity, and are tackled with novel loss functions that drive the extractor towards more separable representations [39,40,58–60]. Effectively training these novel architectures fails with insufficient amount of data. As observed in the previous analysis, FER datasets have a reduced size. Therefore, these limitations opened a new research direction in the context of FER, which is based on network ensembles and on the fusion of different face related tasks (e.g., facial landmark location and face recognition) [61].

In conclusion, the current datasets in the movie domain are still not large enough to allow traditional feature extractors to obtain the desired results. Additionally, physiological variations (such as age, genre, cultural context or levels of expressiveness) and technical inconsistencies (such as people's pose or lighting) are other challenges currently being addressed [61].

3. Proposed Methodology

Based on the evidence discussed in Section 2.2, it becomes clear that there are no sufficient large-scale movie datasets with face-derived emotion annotations. As a direct consequence, there are not many studies that validate the use of FER deep learning models specifically for the movie domain. Therefore, the problem we investigate can be defined through the following research questions: Can current FER datasets and Deep Learning (DL) models for image classification lead to meaningful results? What are the main challenges and limitations of FER in the movie domain? How can current results on affective/emotional analysis with other media be translated to FER in the cinema domain? Are the current emotional models adequate to the cinema domain where expressions are more complex and rehearsed?

Based on these research questions, we defined the following steps as the experimental design:

1. From the list of available datasets provided in Section 2.2.1, we analyzed and selected a dataset for training the DL models and evaluated them in the movie domain;
2. We pre-processed the selected datasets through a facial detector to extract more refined (tightly cropped) facial regions;
3. We tested and benchmarked CNN architectures using accuracy as a performance metric. Furthermore, this first evaluation will also tackle the unbalance of the training dataset;
4. Following the findings reported in Section 2.1 to study an approach for dimensionality reduction which allows to compare our findings with other domains (e.g., audio and text). This final step is divided into two approaches:
   (a) Using only the top-N performing classes;
   (b) Clustering the classes using the emotion clusters found in other studies from the SoA.

Within the datasets introduced in Section 2.2, none fit perfectly the requirements since there is no large-scale FER database in the film domain. Thus, we propose using a cross-database scenario involving two in-the-wild settings that can unite the benefits of a large database with the benefits of a film-based database.

For that purpose, FER2013 [26] was selected based on its size and in the fact that it includes both posed and spontaneous samples. This dataset was created using the Google image search API with 184 different keywords related to emotions, collecting 1000 images for each search query. Images were then cropped in the face region and a face-alignment post-processing phase was conducted. Prior to the experiments, images were grouped by their corresponding emotions. Each image, represented in the 48x48 vector in pixels, is labeled with an encoded emotion.

The number of samples per class of the dataset is presented in Table 3. The imbalance of the dataset is fairly evident, especially between disgust (with only 547 samples) and happy (with 8989 samples) classes. This imbalance is justifiable as it is relatively easy to classify a smile as happiness, while perceiving anger, fear or sadness is a more complicated task for the annotator.

Table 3. The FER2013 number of samples per class.

Emotion	Number of Samples (%)
Angry	4953 (13.8)
Disgust	547 (1.5)
Fear	5121 (14.3)
Happy	8989 (25.0)
Sad	6077 (16.9)
Surprise	4002 (11.1)
Neutral	6198 (17.3)
Total	35,887 (100)

SFEW [41] was also chosen for this analysis since the images were directly collected through film frames. Furthermore, the labels of SFEW are consistent with FER2013 dataset, making the aforementioned cross-database study possible. The original version of the dataset only contained movie stills, while the second version of the dataset comes with pre-processed and aligned faces, and with LPQ (Local Phase Quantization) and PHOG

(Pyramid Histogram of Oriented Gradients) features descriptors used for image feature extraction). Table 4 presents the distribution of the images in the dataset. SFEW was built following a strictly person independent (SPI) protocol, meaning that the train and test datasets do not contain images of the same person.

Table 4. SFEW aligned face samples per class. The test set contains 372 unlabeled images.

	Train	**Validation**	**Test**
Angry	178	77	
Disgust	49	23	
Fear	78	46	
Happy	184	72	
Sad	161	73	
Surprise	94	56	
Neutral	144	84	
Total	888	431	372

4. Results

Following the experimental design referred in Section 3, to set a baseline for our work, we benchmarked several SoA CNN architectures that were initialized with pre-trained weights from ImageNet. The selected backbones were MobileNetV2, Xception, VGG16, VGG19, ResnetV2, InceptionV3 and DenseNet. These models were selected based on their solid performance in other image challenges, with the premise that they could also be applied to FER tasks.

FER2013 was separated into training and testing sets. The baseline models were optimized using cross-entropy and accuracy, for validation purposes, during 25 epochs with a mini batch size of 128. The initial learning rate was set to 0.1, being decreased by a factor of 10% if the validation accuracy did not improve for three epochs. Moreover, the dataset was also extended by applying data augmentation with a probability of 50% on every instance. The selected augmentation methods were horizontal flip and width/height shift (min 10%). Table 5 presents these results for each baseline architecture, while Figures 1–6 illustrate their corresponding confusion matrices.

From the results, it is clear that none of the vanilla models achieved SoA results. In contrast, Xception performed well in inference time, with the second fastest training time of our tests, and the best accuracy result. Taking into account this preliminary analysis, Xception was selected as the baseline model for the purpose of the study to be conducted next.

Since SFEW has few samples, FER2013 was used to train the selected model using a large in-the-wild database of facial expressions. The trained model was tested with SFEW since it contains faces of actors directly extracted from film frames. This enables understanding whether the developed model is robust enough to adapt to a new context. Results are shown in Table 6 and Figures 7 and 8. From the presented numbers, we can conclude that Xception was able to achieve an overall accuracy of 68% in FER2013, which is within state-of-the-art values. Additionally, since FER2013 is a dataset built in a lab environment with a controlled image capturing conditions, these experiments will allow us to analyze whether a network trained in these conditions will have the ability to generalize to the film domain, by testing it with SFEW.

Having achieved the first objective, the next step was to simulate a real testing scenario of the network by submitting it to images taken from films. From a pool of 891 images, results were not satisfactory, reaching an overall accuracy of only 38%. Given this result, the next step was to address an already identified problem: the imbalance of FER2013.

Table 5. Benchmark of CNN architectures.

	Inference Time (ms)	**Accuracy (%)**
MobileNetV2	85.32	52.57
Xception	245.85	64.71
VGG16	1364.90	61.43
ResnetV2	4024.74	62.57
InceptionV3	1127.68	53.14
DenseNet	1131.98	58.86

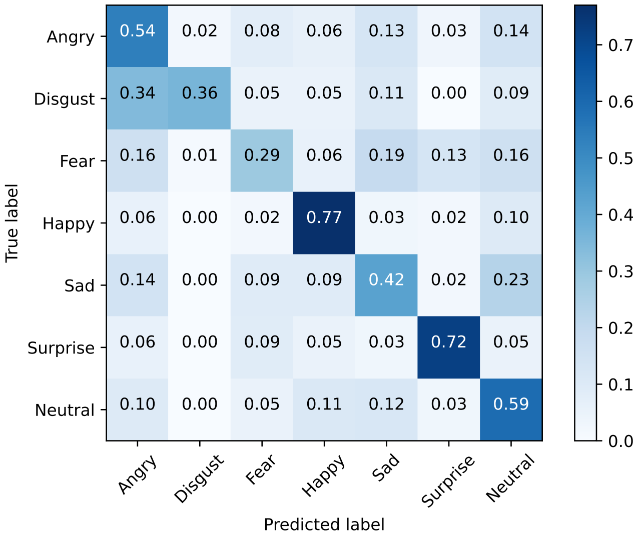

Figure 1. MobileNetV2—training set (FER2013).

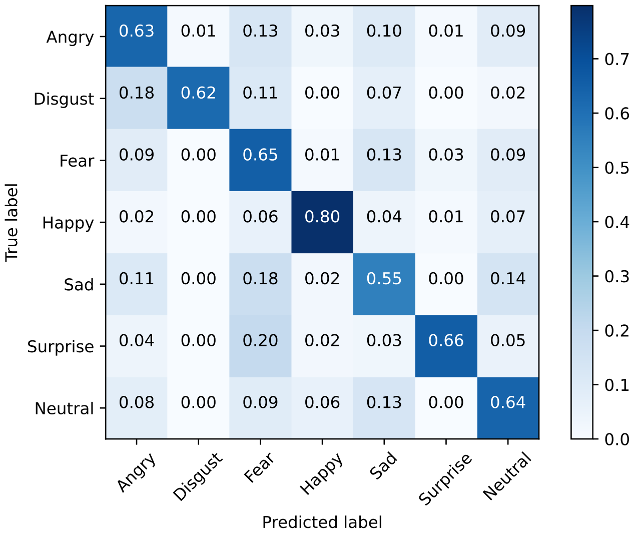

Figure 2. Xception—training set (FER2013).

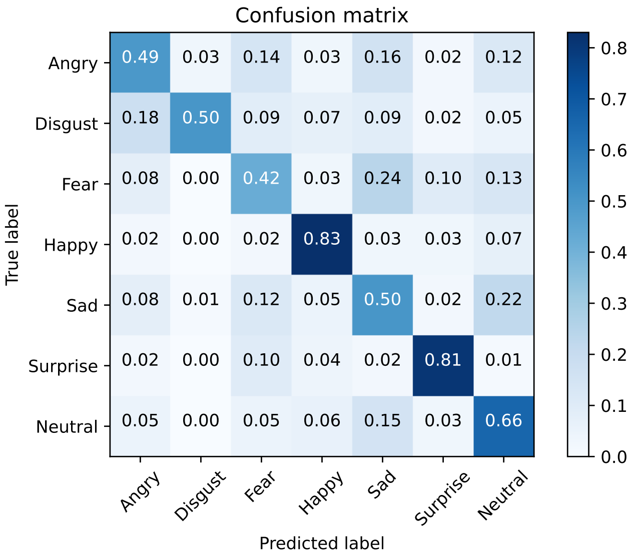

Figure 3. VGG16—training set (FER2013).

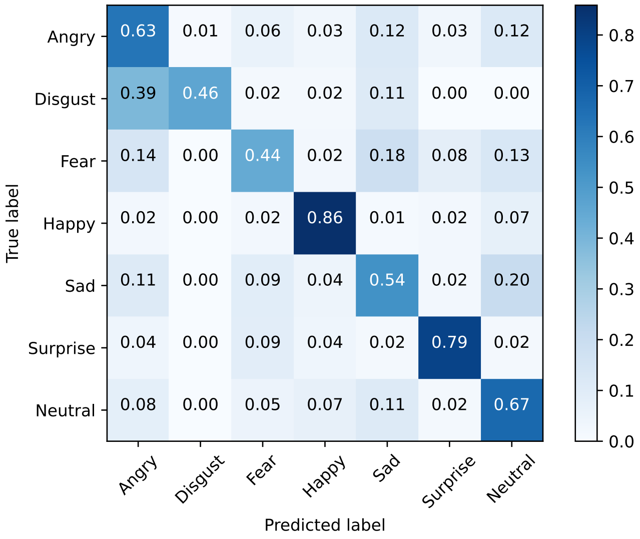

Figure 4. ResNetV2—training set (FER2013).

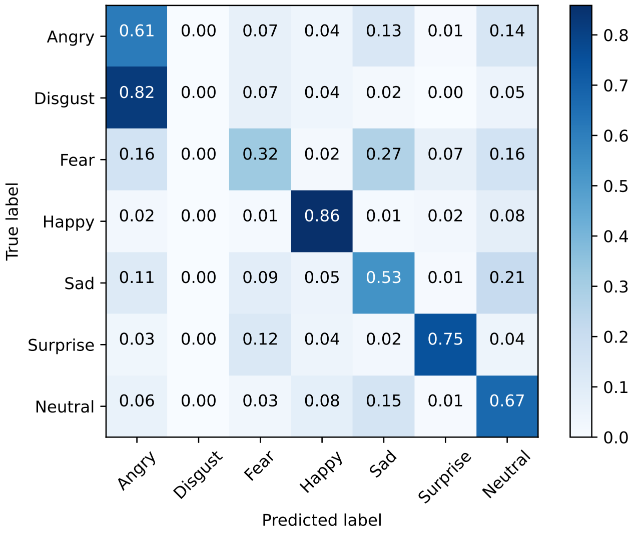

Figure 5. Inceptionv3—training set (FER2013).

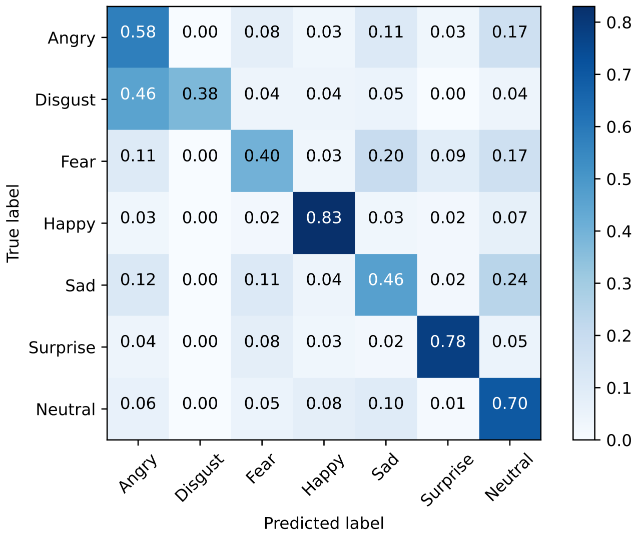

Figure 6. DenseNet—training set (FER2013).

Table 6. Precision, recall and accuracy in FER2013 and SFEW.

	FER2013 (Validation Set)		**SFEW (Test Set)**	
	Prec. (%)	**Rec. (%)**	**Prec. (%)**	**Rec. (%)**
Angry	60	62	53	40
Disgust	56	55	29	10
Fear	58	45	36	31
Happy	87	87	63	82
Sad	60	54	6	1
Surprise	79	80	13	14
Neutral	56	71	22	49
Accuracy	68		38	

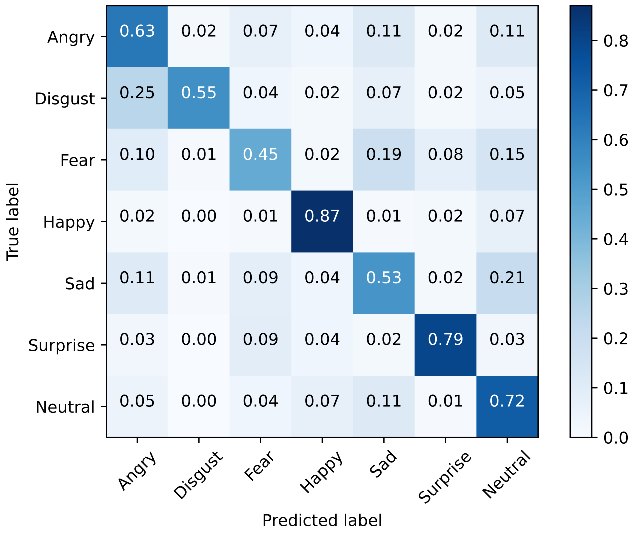

Figure 7. Baseline training set (FER2013).

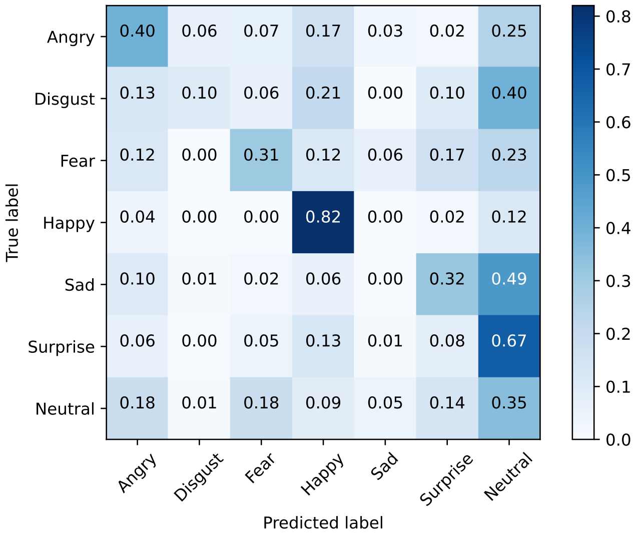

Figure 8. Baseline testing set (SFEW).

4.1. FER2013 Dataset Balancing

To deal with the class imbalance issue, the model was retrained with different class weights which causes the model to "pay more attention" to the examples from an under-represented class. The values used were anger (1.026); disgust (9.407); fear (1.001); happy (0.568); sad (0.849); surprise (1.293); neutral (0.826). Results are illustrated in Table 7.

Table 7. Precision, recall and accuracy in the balanced FER2013 validation set.

	Prec. (%)	**Rec. (%)**
Angry	57	59
Disgust	0	0
Fear	55	38
Happy	83	90
Sad	56	57
Surprise	85	72
Neutral	55	70
Accuracy	66	

Despite the overfit reduction, this approach did not lead to better accuracy results. When tested with SFEW dataset, the obtained results were similar to those already reported.

4.2. Reducing Dimensionality

The gathered evidence in Section 2 and the confusion matrices from the baseline results indicate that there is an overlap of emotions in the affective space. Thus, we propose a reduction in the dimensionality of the problem by reducing the number of emotions to be considered in affective analyses. We demonstrated the effectiveness of this approach firstly by selecting the top-four performing emotions in the previous experiments, and secondly, by selecting the clusters of emotions more clearly demarcated in the studies previously addressed.

4.2.1. Selecting the Top-Four Performing Emotions

The emotions that stood out in the previous tests were happy, surprise, neutral and angry, achieving a accuracy score of 87%, 80%, 71% and 62%, respectively. When training the model solely with these emotions, it was able to achieve an accuracy of 83%, as shown in Table 8. The confusion matrix for this testing scenario is shown in Figure 9.

After analyzing each emotion, we can conclude that by decreasing the size of the problem, the network's performance was improved. When applied to SFEW (Table 8 and Figure 10), the model also demonstrated some improvements with the reduction in dimensionality, going from 38% to 47% accuracy.

Table 8. Precision, recall and accuracy for all the devised methodologies

	Top-4 Performing Emotions				**Clustered Emotions**			
	FER2013		**SFEW**		**FER2013**		**SFEW**	
	Prec. (%)	**Rec. (%)**	**Prec. (%)**	**Rec. (%)**	**Prec. (%)**	**Rec. (%)**	**Prec. (%)**	**Rec. (%)**
Angry	77	74	49	61	-	-	-	-
Disgust	-	-	-	-	-	-	-	-
Fear	-	-	-	-	-	-	-	-
Happy	90	88	96	57	-	-	-	-
Sad	-	-	-	-	-	-	-	-
Surprise	89	87	29	46	-	-	-	-
Neutral	75	81	0	0	77	82	70	52
Positive	-	-	-	-	92	90	93	52
Negative	-	-	-	-	81	78	51	90
Accuracy	83		47		85		64	

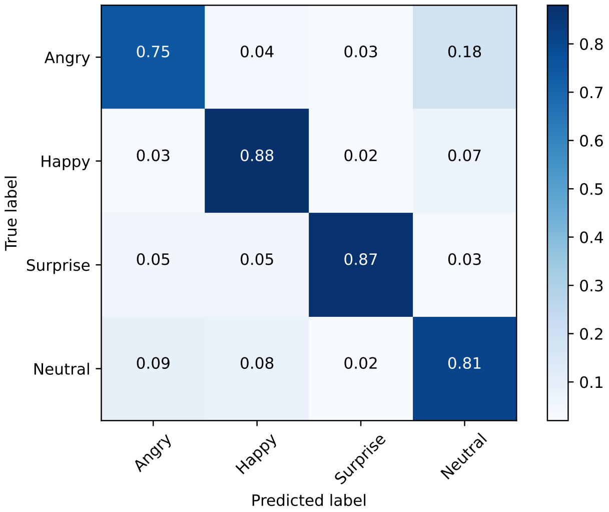

Figure 9. Training set (FER2013) and the top-4 performing emotions.

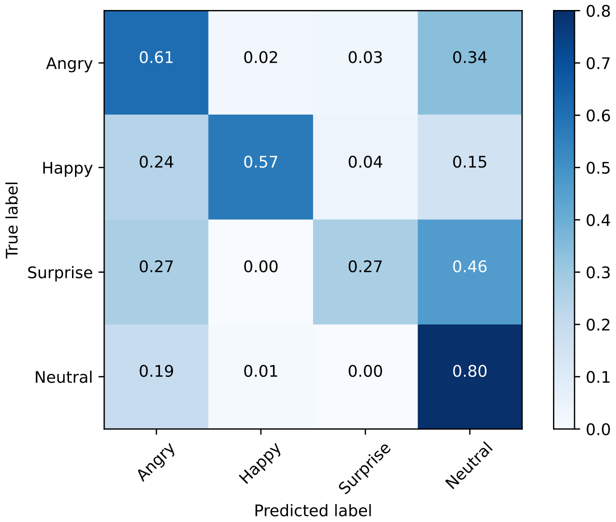

Figure 10. Testing set (SFEW) and the top-4 performing emotions.

4.2.2. Clustered Emotions

Based on the evidence collected in Section 2.1, there are three clearly demarcated emotional clusters: happy (hereafter titled positive), neutral and a third one composed of the angry, sad, fear and disgust (the emotions with a negative connotation—hereafter titled negative). Therefore, another test involving these three clusters was performed. By concentrating only on these three emotions, the network achieved an accuracy of 85%, as illustrated in Table 8. For this methodology, the confusion matrices for the training and testing sets were illustrated, respectively, in Figures 11 and 12.

Testing the "three emotional network" with the SFEW dataset, a score of 64% was achieved, as illustrated in Table 8. Unlike the validation set of FER2013, the emotion with the best performance in SFEW was negative, reaching an accuracy value of 90%.

The best results were obtained when the dimensional reduction took place, so this may be a suitable solution for emotional analysis systems at the cost of losing granularity within the emotions of negative connotation. These results also showed similar emotion clusters as the ones discussed in Section 2 for other domains, that can be depicted in the confusion matrices demonstrated along this section. In particular, they show intersections between the "negative" clusters/classes and the neutral class/cluster (Figures 11 and 12), and within the negative connotation classes (Figures 7 and 8).

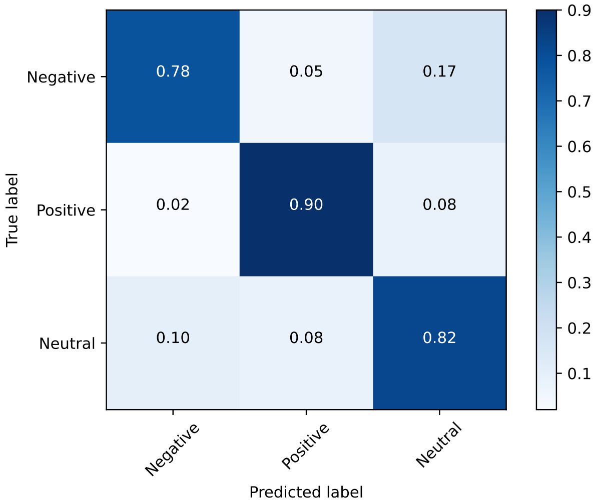

Figure 11. Training set (FER2013) and clustered emotions.

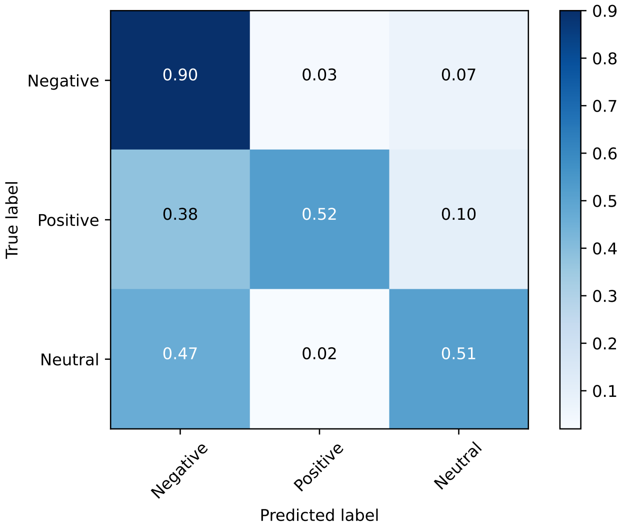

Figure 12. Testing set (SFEW) and clustered emotions.

5. Conclusions

The work described in this paper had as its main objective the definition of an approach for the automatic computation of video-induced emotions using actors' facial expressions. It discusses the main models and theories for representing emotions, discrete and dimensional, along with their respective advantages and limitations. Then, we proceed with the exploration of a theoretical modeling approach from facial expressions to emotions, and discussed a possible approximation between these two very distinctive theories. The contextualization from human and social sciences allowed to foresee that the lack of unanimity in the classification of emotions would naturally have repercussions both in the databases and in the classification models, one of the major bottlenecks of affective analysis.

A systematic validation and benchmark analysis was performed to SoA FER approaches applied to the movie domain. After initial benchmarks, we fine-tuned the chosen model with FER2013, evaluating it with the movie-related dataset, SFEW. During this phase, we noticed several flaws and limitations in these datasets, ranging from class imbalance to even some blank images that do not contain faces. Additionally, we studied, through dimensionality reduction, the hypothesis that clustering observations from the valence–arousal space in other domains are transferable to this approach.

The obtained results show that even if there are still many open challenges related, among others, to the lack of data in the film domain and to the subjectiveness of emotions, the proposed methodology is capable of achieving relevant accuracy standards.

From the work developed and described in this article, several conclusions can be drawn. Firstly, there is a lack of training data both in terms of quantity and quality: there is no publicly available dataset that is large enough for the current deep learning standards. Additionally, within the available databases, there are several inconsistencies in the annotation (using different models of emotion, or even within the same theory of emotion) and image collection processes (illumination variation, occlusions, head-pose variation) that hinder progress in the FER field. Furthermore, the notion of ground truth applied to this context needs to be taken with a grain of salt, since classifying emotion is intrinsically biased in terms of the degree to which it reflects the perception of the emotional experience that the annotator is experiencing.

Paul Ekman's basic emotions model is commonly used in current facial expression classification systems, since it tackles the definition of universal emotions and is widely accepted in the social sciences community. This model was designed by empirical experiences within people from different geographical areas, aiming to understand whether the same facial expressions translate a single emotion, without cultural variations. Hence, Ekman designed seven basic emotions used nowadays in the technological fields, to identify emotion through facial expressions. Current solutions are now quite accurate in this task for a variety of applications, with recent commercial uses, namely in the social networks. However, specifically in the cinema field, analyzing emotions from characters with existing frameworks proved to be an unsatisfying approach. On the one hand, actors are entitled to rehearse a facial expression of a character in a certain context. In this field, emotional representation is acted, thus using Ekman's model might not be a valid solution for the analysis of cinematographic content. For example, by applying current FER approaches to a comedy movie, the results could be flawed because acted emotions in this context should not be translated literally to the exact emotion apparent in the facial expression. In this example, we could obtain a distribution of emotions mostly focused on sadness and surprise, although in the comedy context, the meaning of the character's facial expressions should not be literal: thus, could we consider other basic emotions, with a more complex system that can distinguish an ironic sadness from a real sadness emotion, from a Drama movie? This could be a line of work for future implementations. On the other hand, the images captured in movies are cinematographic, i.e., they are taken in uncontrolled settings, where the environment varies in color, lightness exposure and camera angles. This content variety can be a clear struggle in the classification task, and concretely in the cinema field, it could have a large impact on research results.

Apart from facial expression, there are other characteristics in films that can be used to estimate their emotional charge, as discussed in Section 2. Therefore, as future work, we expect to use facial landmarks to obtain facial masks and, alongside the original image, use them as input to the model. This information might be leveraged as embedded regularization to weight faces' information in the classification of the conveyed emotions of movies. Furthermore, temporal information regarding the evolution of visual features might also be worth exploring since they are commonly used to convey emotions in cinematographic pieces. Regarding the annotation subjectiveness, we also considered that designing intuitive user interfaces that enable the annotator to perceive the differences between discrete emotion classes is also a future path to enhance the annotation process and quality, and to reduce the amount of noise in the construction of new datasets for the field.

Author Contributions: Conceptualization, J.A., L.V., I.N.T. and P.V.; methodology, J.A., L.V., I.N.T. and P.V.; software, J.A; validation, J.A., L.V., I.N.T. and P.V.; formal analysis, J.A., L.V., I.N.T. and P.V.; investigation, J.A., L.V., I.N.T. and P.V.; resources, J.A., L.V., I.N.T. and P.V.; data curation, J.A., L.V., I.N.T. and P.V.; writing—original draft preparation, J.A., L.V., I.N.T. and P.V.; writing—review and editing, J.A., L.V., I.N.T. and P.V.; visualization, J.A., L.V., I.N.T. and P.V.; supervision, L.V., I.N.T. and P.V.; project administration, L.V., I.N.T. and P.V.; funding acquisition, P.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially financed by the ERDF—European Regional Development Fund—through the Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement and through the Portuguese National Innovation Agency (ANI) as a part of project CHIC: NORTE-01-0247-FEDER-0224498; and by National Funds through the Portuguese funding agency, FCT—Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Ekman, P.; Keltner, D. Universal facial expressions of emotion. In *Nonverbal Communication: Where Nature Meets Culture*; Segerstrale, U., Molnar, P., Eds.; Routledge: London, UK, 1997; pp. 27–46.
2. Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
3. Ortony, A.; Clore, G.L.; Collins, A. *The Cognitive Structure of Emotions*; Cambridge University Press: Cambridge, UK, 1990.
4. Prinz, J.J. *Gut Reactions: A Perceptual Theory of Emotion*; Oxford University Press: Oxford, UK, 2004.
5. Parrott, W.G. *Emotions in Social Psychology: Essential Readings*; Psychology Press: Philadelphia, PA, USA, 2001.
6. Friesen, E.; Ekman, P. Facial action coding system: A technique for the measurement of facial movement. *Palo Alto* **1978**, *3*, 5.
7. Ekman, P.; Rosenberg, E.L. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*; Oxford University Press: Oxford, UK, 2020.
8. Fabian Benitez-Quiroz, C.; Srinivasan, R.; Martinez, A.M. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5562–5570.
9. Posner, J.; Russell, J.A.; Peterson, B.S. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol.* **2005**, *17*, 715. [CrossRef] [PubMed]
10. Cacioppo, J.T.; Berntson, G.G.; Larsen, J.T.; Poehlmann, K.M.; Ito, T.A. The psychophysiology of emotion. In *Handbook of Emotions*; Guilford Press: New York, NY, USA, 2000; Volume 2, pp. 173–191.
11. Jack, R.E.; Garrod, O.G.; Yu, H.; Caldara, R.; Schyns, P.G. Facial expressions of emotion are not culturally universal. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 7241–7244. [CrossRef]
12. Saarni, C. *The Development of Emotional Competence*; Guilford Press: New York, NY, USA, 1999.
13. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161. [CrossRef]
14. Whissell, C.M. The dictionary of affect in language. In *The Measurement of Emotions*; Elsevier: Amsterdam, The Netherlands, 1989; pp. 113–131.

15. Mehrabian, A. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Curr. Psychol.* **1996**, *14*, 261–292. [CrossRef]
16. Greenwald, M.K.; Cook, E.W.; Lang, P.J. Affective judgment and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli. *J. Psychophysiol.* **1989**. Available online: https://psycnet.apa.org/record/1990-03841-001 (accessed on 18 June 2021).
17. Fontaine, J.R.; Scherer, K.R.; Roesch, E.B.; Ellsworth, P.C. The world of emotions is not two-dimensional. *Psychol. Sci.* **2007**, *18*, 1050–1057. [CrossRef] [PubMed]
18. Gebhard, P. ALMA: A layered model of affect. In Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems, Utrecht, The Netherlands, 25–29 July 2005; pp. 29–36.
19. Shi, Z.; Wei, J.; Wang, Z.; Tu, J.; Zhang, Q. Affective transfer computing model based on attenuation emotion mechanism. *J. MultiModal User Interfaces* **2012**, *5*, 3–18. [CrossRef]
20. Landowska, A. Towards new mappings between emotion representation models. *Appl. Sci.* **2018**, *8*, 274. [CrossRef]
21. Bradley, M.M.; Lang, P.J. *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings*; Technical report, Technical report C-1; The Center for Research in Psychophysiology, University of Florida: Gainesville, FL, USA, 1999.
22. Krcadinac, U.; Pasquier, P.; Jovanovic, J.; Devedzic, V. Synesketch: An open source library for sentence-based emotion recognition. *IEEE Trans. Affect. Comput.* **2013**, *4*, 312–325. [CrossRef]
23. Riegel, M.; Wierzba, M.; Wypych, M.; Żurawski, Ł.; Jednoróg, K.; Grabowska, A.; Marchewka, A. Nencki affective word list (NAWL): The cultural adaptation of the Berlin affective word list–reloaded (BAWL-R) for Polish. *Behav. Res. Methods* **2015**, *47*, 1222–1236. [CrossRef]
24. Wierzba, M.; Riegel, M.; Wypych, M.; Jednoróg, K.; Turnau, P.; Grabowska, A.; Marchewka, A. Basic emotions in the Nencki Affective Word List (NAWL BE): New method of classifying emotional stimuli. *PLoS ONE* **2015**, *10*, e0132305. [CrossRef]
25. Eerola, T.; Vuoskoski, J.K. A comparison of the discrete and dimensional models of emotion in music. *Psychol. Music.* **2011**, *39*, 18–49. [CrossRef]
26. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H.; et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*; Springer: Berlin, Heidelberg, Germany, 2013; pp. 117–124.
27. Dhall, A.; Goecke, R.; Joshi, J.; Hoey, J.; Gedeon, T. Emotiw 2016: Video and group-level emotion recognition challenges. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 427–432.
28. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. Collecting large, richly annotated facial-expression databases from movies. *IEEE Ann. Hist. Comput.* **2012**, *19*, 34–41. [CrossRef]
29. Kossaifi, J.; Tzimiropoulos, G.; Todorovic, S.; Pantic, M. AFEW-VA database for valence and arousal estimation in-the-wild. *Image Vis. Comput.* **2017**, *65*, 23–36. [CrossRef]
30. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [CrossRef]
31. Kollias, D.; Zafeiriou, S. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv* **2018**, arXiv:1811.07770.
32. McDuff, D.; Amr, M.; El Kaliouby, R. Am-fed+: An extended dataset of naturalistic facial expressions collected in everyday settings. *IEEE Trans. Affect. Comput.* **2018**, *10*, 7–17. [CrossRef]
33. 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.
34. Lyons, M.; Akamatsu, S.; Kamachi, M.; Gyoba, J. Coding facial expressions with gabor wavelets. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 200–205.
35. Calvo, M.G.; Lundqvist, D. Facial expressions of emotion (KDEF): Identification under different display-duration conditions. *Behav. Res. Methods* **2008**, *40*, 109–115. [CrossRef]
36. 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. [CrossRef]
37. Valstar, M.; Pantic, M. Induced disgust, happiness and surprise: An addition to the mmi facial expression database. In Proceedings of the 3rd International Workshop on EMOTION (Satellite of LREC): Corpora for Research on Emotion and Affect, Valletta, Malta, 17–23 May 2010; p. 65
38. Zhao, G.; Huang, X.; Taini, M.; Li, S.Z.; PietikäInen, M. Facial expression recognition from near-infrared videos. *Image Vis. Comput.* **2011**, *29*, 607–619. [CrossRef]
39. Li, S.; Deng, W.; Du, J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2852–2861.
40. Li, S.; Deng, W. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Trans. Image Process.* **2018**, *28*, 356–370. [CrossRef]
41. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. *Acted Facial Expressions in the Wild Database*; Technical Report TR-CS-11; Australian National University: Canberra, Australia, 2011; Volume 2, p. 1.

42. Cohn, J.F.; Ertugrul, I.O.; Chu, W.S.; Girard, J.M.; Jeni, L.A.; Hammal, Z. Affective facial computing: Generalizability across domains. In *Multimodal Behavior Analysis in the Wild*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 407–441.
43. Meng, D.; Peng, X.; Wang, K.; Qiao, Y. Frame attention networks for facial expression recognition in videos. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3866–3870.
44. Zhang, Z.; Luo, P.; Loy, C.C.; Tang, X. From facial expression recognition to interpersonal relation prediction. *Int. J. Comput. Vis.* **2018**, *126*, 550–569. [CrossRef]
45. Breuer, R.; Kimmel, R. A deep learning perspective on the origin of facial expressions. *arXiv* **2017**, arXiv:1705.01842.
46. Pramerdorfer, C.; Kampel, M. Facial expression recognition using convolutional neural networks: State of the art. *arXiv* **2016**, arXiv:1612.02903.
47. Kim, D.H.; Baddar, W.J.; Jang, J.; Ro, Y.M. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Trans. Affect. Comput.* **2017**, *10*, 223–236. [CrossRef]
48. Hamester, D.; Barros, P.; Wermter, S. Face expression recognition with a 2-channel convolutional neural network. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12–17 July 2015; pp. 1–8.
49. Minaee, S.; Abdolrashidi, A. Deep-emotion: Facial expression recognition using attentional convolutional network. *arXiv* **2019**, arXiv:1902.01019.
50. Yu, Z.; Zhang, C. Image based static facial expression recognition with multiple deep network learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 435–442.
51. Kim, B.K.; Lee, H.; Roh, J.; Lee, S.Y. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 427–434.
52. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Trans. Image Process.* **2018**, *28*, 2439–2450. [CrossRef] [PubMed]
53. Yang, H.; Zhang, Z.; Yin, L. Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 294–301.
54. Ng, H.W.; Nguyen, V.D.; Vonikakis, V.; Winkler, S. Deep learning for emotion recognition on small datasets using transfer learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 443–449.
55. Ding, H.; Zhou, S.K.; Chellappa, R. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 118–126.
56. Yao, A.; Cai, D.; Hu, P.; Wang, S.; Sha, L.; Chen, Y. HoloNet: Towards robust emotion recognition in the wild. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 472–478.
57. Hu, P.; Cai, D.; Wang, S.; Yao, A.; Chen, Y. Learning supervised scoring ensemble for emotion recognition in the wild. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; pp. 553–560.
58. 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 & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 302–309.
59. Guo, Y.; Tao, D.; Yu, J.; Xiong, H.; Li, Y.; Tao, D. Deep neural networks with relativity learning for facial expression recognition. In Proceedings of the 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Seattle, WA, USA, 11–15 July 2016; pp. 1–6.
60. Liu, X.; Vijaya Kumar, B.; You, J.; Jia, P. Adaptive deep metric learning for identity-aware facial expression recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 20–29.
61. Li, S.; Deng, W. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2020**. [CrossRef]