

Review

Facial Expression Recognition Using Computer Vision: A Systematic Review

Daniel Canedo *  and António J. R. Neves 

IEETA/DETI, University of Aveiro, 3810-193 Aveiro, Portugal; an@ua.pt

* Correspondence: danielduartecanedo@ua.pt; an@ua.pt

Received: 1 October 2019; Accepted: 29 October 2019; Published: 2 November 2019



Abstract: Emotion recognition has attracted major attention in numerous fields because of its relevant applications in the contemporary world: marketing, psychology, surveillance, and entertainment are some examples. It is possible to recognize an emotion through several ways; however, this paper focuses on facial expressions, presenting a systematic review on the matter. In addition, 112 papers published in ACM, IEEE, BASE and Springer between January 2006 and April 2019 regarding this topic were extensively reviewed. Their most used methods and algorithms will be firstly introduced and summarized for a better understanding, such as face detection, smoothing, Principal Component Analysis (PCA), Local Binary Patterns (LBP), Optical Flow (OF), Gabor filters, among others. This review identified a clear difficulty in translating the high facial expression recognition (FER) accuracy in controlled environments to uncontrolled and pose-variant environments. The future efforts in the FER field should be put into multimodal systems that are robust enough to face the adversities of real world scenarios. A thorough analysis on the research done on FER in Computer Vision based on the selected papers is presented. This review aims to not only become a reference for future research on emotion recognition, but also to provide an overview of the work done in this topic for potential readers.

Keywords: facial expression recognition; emotion recognition; computer vision; machine learning; action units; deep learning; facial features; review article

1. Introduction

Emotion recognition is being actively explored in Computer Vision research. With the recent rise and popularization of Machine Learning [1] and Deep Learning [2] techniques, the potential to build intelligent systems that accurately recognize emotions became a closer reality. However, this problem is shown to be more and more complex with the progress of fields that are directly linked with emotion recognition, such as psychology and neurology. Micro-expressions, electroencephalography (EEG) signals, gestures, tone of voice, facial expressions, and surrounding context are some terms that have a powerful impact when identifying emotions in a human [3]. When all of these variables are pieced together with the limitations and problems of the current Computer Vision algorithms, emotion recognition can get highly complex.

Facial expressions are the main focus of this systematic review. Generally, an FER system consists of the following steps: image acquisition, pre-processing, feature extraction, classification, or regression, as shown in Figure 1.

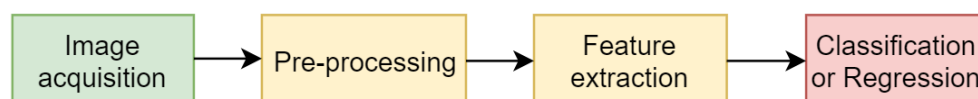


Figure 1. Conventional FER system diagram.

To be able to get a proper facial expression classification, it is highly desirable to provide the most relevant data to the classifier, in the best possible conditions. In order to do that, a conventional FER system will firstly pre-process the input image. One pre-processing step that is common among most reviewed papers is face detection.

Face detection techniques are able to create bounding boxes that delimit detected faces, which are the desired regions of interest (ROIs) for a conventional FER system. This task is still challenging, and it is not guaranteed that all faces are going to be detected in a given input image. This is especially true when acquiring images from an uncontrolled environment, where there may be movement, harsh lighting conditions, different poses, great distances, among other factors [4].

When the faces are properly detected, a conventional FER system will process the retrieved ROIs in order to prepare the data that will be fed into the classifier. Normally, this pre-processing step is divided into several substeps, such as intensity normalization for illumination changes, noise filters for image smoothing, data augmentation (DA) [5] to increase the training data, rotation correction for the rotated faces, image resizing for different ROI sizes, image cropping for a better background filtering, among others. After the pre-processing, one can retrieve relevant features from the pre-processed ROIs. There are numerous features that can be selected, such as Actions Units (AUs) [6], motion of certain facial landmarks, distance between facial landmarks, facial texture, gradient features, and so forth. Then, these features are fed into a classifier. Generally, the classifiers used in an FER system are Support Machine Vectors (SVMs) [7] or Convolutional Neural Networks (CNNs) [8].

This systematic review is organized as follows: Section 2 presents the paper selection criteria for this systematic review; Section 3 presents the most popular FER databases; Section 4 presents the most popular pre-processing methods in FER; Section 5 presents the most popular feature extraction methods in FER; Section 6 presents the most popular classifiers in FER; Section 7 presents the most relevant results obtained by the selected works, as well as a discussion; Section 8 presents insights on the Emotion Recognition in the Wild Challenge (EmotiW); Section 9 presents the conclusion and last remarks of this systematic review.

2. Selection Criteria

It was decided to search the literature from January 2006 (when the publications on machine learning in emotion recognition started emerging) to April 2019. The used databases were ACM [9], IEEE [10], BASE [11], and Springer [12]. These databases were chosen because of their relevance in Computer Vision.

The searching strategy consisted of Open Access articles, journals, conference objects, and manuscripts. There are several keywords and variants that cover the FER field; therefore, it was decided to limit the searched keywords to the most popular two within this topic: “Emotion Recognition” or “Facial Expression Recognition”. The full text of the works was considered during the searching process. The results were the following:

- ACM: 239 papers,
- IEEE: 118 papers,
- BASE: 78 papers,
- Springer: 369 papers.

All the papers were assessed, resulting in a total of 783 non-duplicate papers. The selection criteria for this systematic review only include works that explore FER using Computer Vision. First, all the titles and abstracts were quickly appraised in order to exclude studies that were not relevant to the scope of this systematic review or did not meet the selection criteria, resulting in 185 papers. Second, the full text of the remaining papers was carefully evaluated in order to filter out the remaining studies that did not meet the selection criteria, resulting in 112 papers. The following categories of papers were excluded:

1. Theoretical studies,

2. Studies that are not related with Computer Vision,
3. Surveys and Thesis,
4. Dataset publications,
5. Older iterations of the same studies.

Figure 2 sums up the search and exclusion steps of this systematic review. It is important to mention that some included papers did not present results, they were not conclusive or their procedures and/or results deviate from the mainstream scope; therefore, they were not included in Section 7 [13–49]. However, those papers still contributed to the remaining sections.

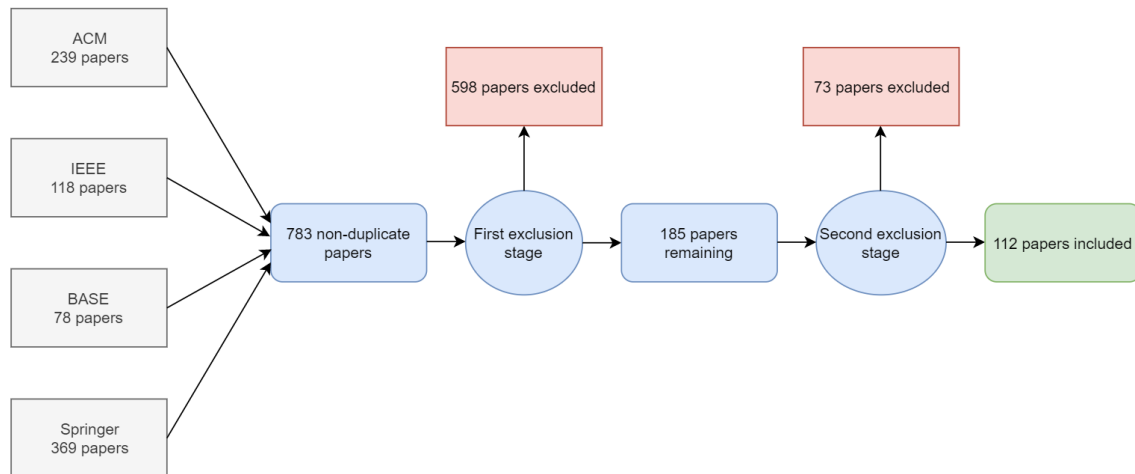


Figure 2. Flowchart of the paper selection for this systematic review.

Figure 3 shows how many papers were found for each year in the searched time frame, displaying a clear increase of interest in emotion recognition. It only appears to decrease in 2019 because, as mentioned above, the searched time frame goes from January 2006 to April 2019.

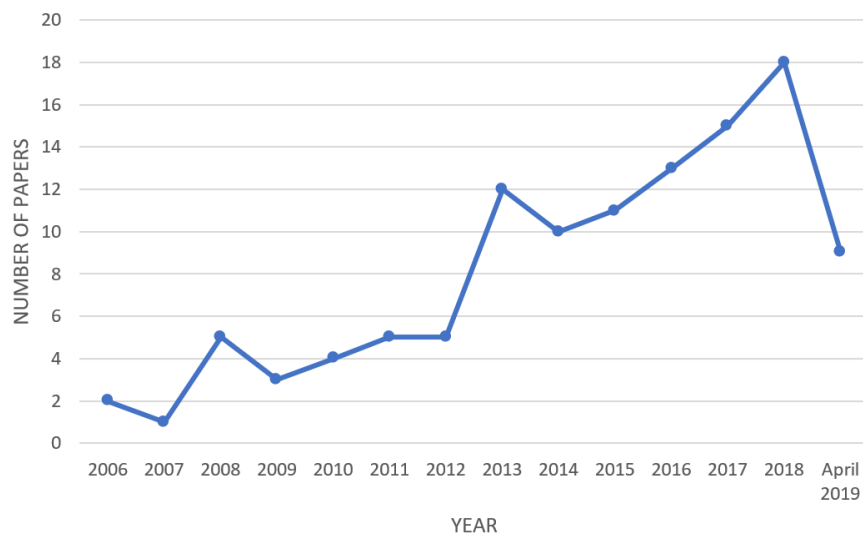


Figure 3. Number of papers found for each year during the paper selection stage.

3. FER Databases

In order to build FER systems that are able to obtain results which can further be compared with related works, researchers working in the FER field have numerous databases at their disposal. Most databases are built on 2D static images or 2D video sequences; however, there are some databases that contain 3D images. An FER system built on a 2D approach has the limitation of handling different

poses poorly, since most 2D databases only contain frontal faces. A 3D approach is potentially capable of handling the pose variation problem. Most FER databases are labeled with the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise), plus the neutral expression. Some FER databases are built on controlled environments (generally inside a laboratory with controlled lighting conditions), while others are built on uncontrolled or wild environments. Furthermore, the subjects of some FER databases were asked to pose certain emotions towards a reference, while others tried to stimulate spontaneous and genuine facial expressions. This section introduces the most popular databases used in the reviewed works:

- **The Extended Cohn–Kanade database (CK+)** [50]: contains 593 image sequences of posed and non-posed emotions. In addition, 123 participants were 18 to 50 years of age, 69% female, 81%, Euro-American, 13% Afro-American, and 6% other groups. The images were digitized into either 640×490 or 640×480 resolution and are mostly gray. Each sequence was built on frontal views and 30-degree views, starting with a neutral expression up until the peak emotion (last frame of the sequence). Most sequences are labeled with eight emotions: anger, disgust, contempt, fear, neutral, happiness, sadness, and surprise.
- **The Japanese Female Facial Expression database (JAFFE)** [51]: contains 213 images of six basic emotions, plus the neutral expression posed by 10 Japanese female models. Each image has been labeled by 60 Japanese subjects.
- **Binghamton University 3D Facial Expression database (BU-3DFE)** [52]: contains 606 3D facial expression sequences captured from 101 subjects. The texture video has a resolution of about 1040×1329 pixels per frame. The resulting database consists of 58 female and 43 male subjects, with a large variety of ethnic/racial ancestries. This database was built on the six basic emotions, plus the neutral expression.
- **Facial Expression Recognition 2013 database (FER-2013)** [53]: was created using the Google image search API to search for images of faces that match a set of 184 emotion-related keywords like “blissful”, “enraged”, etc. These keywords were combined with words related to gender, age or ethnicity, leading to 35,887 grayscale images with a 48×48 resolution, mapped into the six basic emotions, plus the neutral expression.
- **Emotion Recognition in the Wild database (EmotiW)** [54]: contains two sub-databases, Acted Facial Expression in the Wild (AFEW) and the Static Facial Expression in the Wild (SFEW). AFEW contains videos (image sequences including audio) and SFEW contains static images. This database was built on the six basic emotions, plus the neutral expression and the image size is 128×128 .
- **MMI database** [55]: contains over 2900 videos and high-resolution still images of 75 subjects. It is fully annotated for the presence of AUs in videos, and partially coded on frame-level, indicating for each frame whether an AU is in either the neutral, onset, apex or offset phase. This database was built on six emotions: anger, disgust, fear, happiness, sadness, and surprise.
- **eNTERFACE’05 Audiovisual Emotion database** [56]: contains 42 subjects from 14 different nationalities. Among the 42 subjects, 81% were men and the remaining 19% were women. In addition, 31% of the subjects wore glasses, while 17% had a beard, which consists of video sequences (including audio) with a 720×576 resolution. This database was built on six emotions: anger, disgust, fear, happiness, sadness, and surprise.
- **Karolinska Directed Emotional Faces database (KDEF)** [57]: contains a set of 4900 pictures of human facial expressions. The set contains 70 individuals (35 females and 35 males) displaying the six basic emotions, plus the neutral expression. Each expression is viewed from five different angles and was photographed in two sessions.
- **Radboud Faces Database (RaFD)** [58]: contains a set of pictures of 67 models (including Caucasian males and females and Moroccan Dutch males) displaying eight emotional expressions (anger, disgust, contempt, fear, neutral, happiness, sadness, and surprise), amounting to 120 images per model. Each emotion is shown with three different gaze directions and all pictures were taken from five camera angles simultaneously. The image size is 1024×681 .

Figure 4 shows some examples of the introduced FER databases:

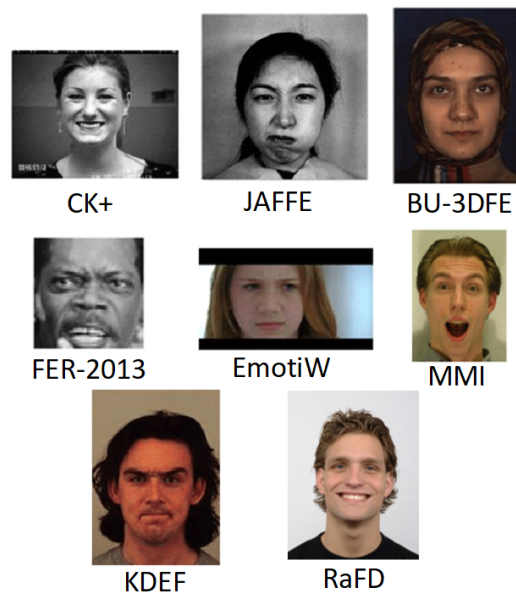


Figure 4. Examples of the introduced FER databases [50–58].

Table 1 sums up these introduced databases and provides their website.

Table 1. Summarized FER databases.

Database	Capacity	Emotion	Environment	Facial Expressions	Website
CK+	593 videos	Posed	Controlled	8	[59]
JAFFE	213 images	Posed	Controlled	7	[60]
BU-3DFE	606 videos	Posed and Spontaneous	Controlled	7	[61]
FER-2013	35,887 images	Posed and Spontaneous	Uncontrolled	7	[62]
EmotiW	1268 videos and 700 images	Spontaneous	Uncontrolled	7	[63]
MMI	2900 videos	Posed	Controlled	6	[64]
eNTERFACE'05	1166 videos	Spontaneous	Controlled	6	[65]
KDEF	4900 images	Posed	Controlled	7	[66]
RaFD	8040 images	Posed	Controlled	8	[67]

4. Pre-Processing

Pre-processing is one of the most important phases, not only in FER systems, but in any Machine Learning based system. Analyzing raw data without any kind of screening can lead to unwanted results. This is why assuring the quality of data before extracting its relevant features is vital. The following subsections present the most popular and effective pre-processing techniques in the reviewed works.

4.1. Face Detection

Generally, face detection is the very beginning of an FER system. This technique is responsible for selecting the ROI of an input image that will be fed to the next steps of the FER system. Most reviewed papers used the classic Viola–Jones face detector [68] from 2004.

The Viola–Jones face detector is a Machine Learning based approach where a cascade function is trained from a lot of positive images (images with faces) and negative images (images without faces). Haar features are used in this algorithm and they are applied in all training images to find the best threshold which will classify the faces as positive or negative detections. Figure 5 shows how these

features are selected by AdaBoost (learning algorithm that selects a small number of critical visual features from a very large set of potential features).

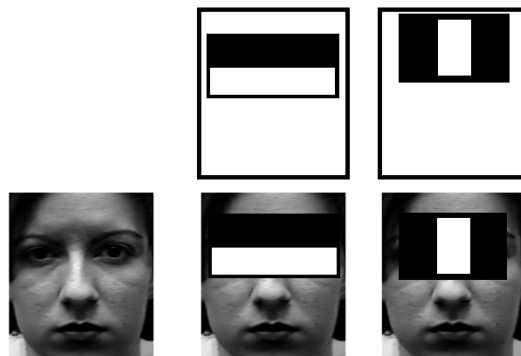


Figure 5. Features selected by the AdaBoost learning algorithm in a face from the CK+ database.

Two examples of features are shown in the top row and then overlaid on a training face in the bottom row. Basically, these features measure the difference in intensity between the white region and the black region. Typically, one region is darker than the other, and that is how the features are selected to detect faces.

Very few works used other face detectors, such as the Dlib library [69] and Multi-task Cascade Convolution Neural Network (MTCNN) [70]. The Dlib face detector uses an ensemble of regression trees to regress the location of 68 facial landmarks from a sparse subset of intensity values extracted from an input image and, consequently, detect where the faces are. Figure 6 shows a detected face from the CK+ database and its 68 facial landmarks using the Dlib library.

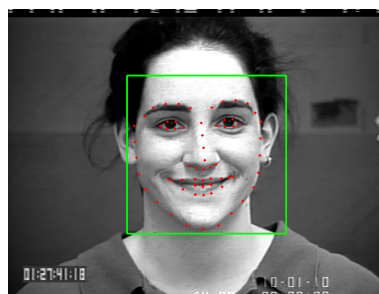


Figure 6. Detected face from the CK+ database using the Dlib library.

As for MTCNN, it goes through three stages to output a proper face detection, and, in each stage, the face that is being analyzed goes through a CNN. The first stage obtains the candidate windows and their bounding box regression vectors, merging highly overlapped candidates [71]. The second stage feeds those candidates to another CNN, which rejects a large number of false candidates. The third stage is similar to the second one, but it also outputs five facial landmarks' positions. Figure 7 shows a detected face from the CK+ database and its five facial landmarks using MTCNN.



Figure 7. Detected face from the CK+ database using MTCNN.

Despite Viola–Jones being the most used face detector choice in the reviewed papers, MTCNN outperforms it across several challenging benchmarks for face detection and face alignment, while keeping real-time performance. The main reason Viola–Jones was the most used face detector is that most reviewed works tested their FER systems in controlled environment databases, where Viola–Jones can smoothly detect the faces. Works that tested their FER systems in uncontrolled environment databases generally used MTCNN.

4.2. Geometric Transformations

Even if the faces are detected in an input image, it does not mean that they are in proper conditions to be analyzed. Some problems that can arise from these detected ROIs are rotation, scale, and noise [72]. One has to guarantee that the face to be classified is as geometrically similar as possible to the faces used when training the classifier. That way, the classifier is apter to produce more trustworthy results.

Some reviewed works applied geometric transformations to faces that were not detected in the best conditions. Starting with the rotation, the most popular way to correct it is by using the facial landmarks outputted by the face detector. Typically, the reviewed works considered two facial landmarks that form an angle of zero in the horizontal axis, when a face is aligned. To perform face alignment of a rotated face, a rotation transformation is applied to align those two facial landmarks with the horizontal axis, until the angle formed by them is zero. Figure 8 shows a rotation correction on a face from the CK+ database.

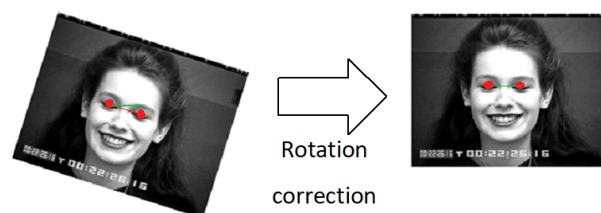


Figure 8. Rotation correction on a face from the CK+ database.

As for the scale problem, it is because of the different distances in which faces can be detected: closer detected faces have bigger ROIs than farther ones. Since it is crucial to feed the next stages of the FER system with the same size of ROIs, the technique used in the reviewed works is resizing every ROI to a predetermined size (spatial normalization).

The noise problem is mainly the background in the detected ROIs. It is important to remove the background from the original ROIs since it can decrease the accuracy of the classifier by adding one more variable to the problem: distinguishing between foreground and background. Most works seemed to overlook this kind of noise, but there are a few that tried to crop the ROI even more in order to filter the background. The most popular approach to do this is to use the facial landmarks and the bounding box outputted by the face detector. By calculating the distances between relevant facial landmarks, it is possible to reduce the ROI dimension obtained from the face detector and filter the background noise. Figure 9 shows an approach of background removal on a face from the CK+ database.



Figure 9. Background removal using the distance between eyes of a face from the CK+ database.

4.3. Image Processing

Having a proper geometric transformed ROI might not be enough to prepare image data. There are several image processing techniques to accentuate relevant features that are going to be used in the classifier, but most FER systems of the reviewed papers used the following ones:

4.3.1. Smoothing

Smoothing in image processing is often necessary. By smoothing an image, one can capture relevant patterns while filtering the noise. That way, smoothing can provide robustness to the data that is going to be analyzed. There are several ways to smooth an image, but the most popular ones in the reviewed papers are through a bilateral filter [73] or a Gaussian filter [74].

A bilateral filter is effective in noise removal while keeping edges sharp, since it uses a gaussian function of space for smoothing only the nearby pixels and a gaussian function of intensity for smoothing pixels that have a similar intensity to the central pixel. That way, it can preserve the edges, since they usually have high intensity variations.

A Gaussian filter is effective in removing Gaussian noise from an image. It takes the neighborhood around the central pixel and find its Gaussian weighted average. This Gaussian filter is a function of space alone; therefore, it will also smooth the edges.

In a conventional FER system, a bilateral filter has the upper hand when smoothing a face, since it keeps the edges from being blurred. Edges in a face normally delimit facial features such as eyes, nose, eyebrows and mouth. Smoothing these critical features for FER might be something undesirable.

4.3.2. Histogram Equalization

Histograms plot the intensity distribution of an image. Because of potential different lighting conditions that extracted faces may present, the recognition accuracy will likely suffer. There are histogram based algorithms in Computer Vision that are able to stretch out this distribution, meaning that overexposed or underexposed regions of the face will have their intensity uniformed. This method has the advantage of improving the contrast in an image, highlighting facial features and reducing the interference caused by different lighting conditions. However, it may also increase the contrast of background noise [75]. Reviewed works that used this pre-processing approach tended to explore Histogram equalization (HE) [76]. Figure 10 shows a few results on the CK+ database using HE.



Figure 10. Results in some faces of the CK+ database using HE.

4.3.3. Data Augmentation

Emotion recognition databases are generally small, which is something undesirable for Machine Learning classifiers. Training on small data can lead to overfitting [77], which is a common problem in Machine Learning models. A model overfits when it classifies accurately data used for training, but its accuracy drops considerably when classifying data outside the training set (poor generalization). Therefore, overfitting can be spotted when training the model: the accuracy on the training data are

high, but the accuracy on the validation data is significantly lower. One way to deal with this problem that is often a consequence of using small databases for training is through DA.

DA is a technique that enables the increase of data by modifying it reasonably. These modifications can be cropping, flipping, rotating, zooming, rescaling, brightness changing, shifting, among others. It is important to keep track of these modifications; otherwise, this method might change the proper meaning of the training samples and confuse the classifier. Figure 11 shows an example of DA.



Figure 11. DA results on a face from the CK+ database.

4.3.4. Principal Component Analysis

PCA [78] is a method used to reduce the dimension of a large number of features keeping most of their information. At the potential expense of a small amount of accuracy, one can simplify data by reducing the huge amount of variables. In an FER system, this algorithm can be used to reduce redundant facial features, leading to an increase of the computational efficiency.

5. Feature Extraction

Once the pre-processing phase is over, one can extract the relevant highlighted features. In a conventional FER system, the relevant features are facial features. The quality of these features plays a huge role in system accuracy; therefore, several techniques to extract features were developed in Computer Vision. The most popular feature extraction techniques in the reviewed works are the following:

5.1. Local Binary Patterns

LBP [79] is known as one of the best methods for texture processing. This algorithm aims to compare a center pixel with its 3×3 square neighborhood. If the neighbor pixel value is greater than or equal to the center pixel value, then it takes the value "1", else it takes the value "0". Afterwards, one can take the resulting binary code from this operation and set the center pixel with its decimal value. Figure 12 shows an example of this process.

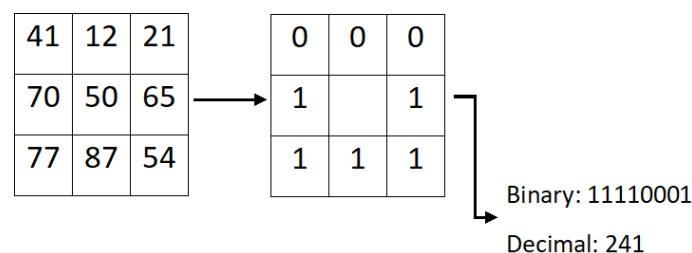


Figure 12. LBP operation.

In FER systems, LBP has the potential to highlight relevant facial features for emotion recognition, such as eyebrows, eyes, nose, and mouth. However, it has the disadvantage of being sensitive to noise: since it is a method based on intensity differences, it is affected by image noise when processing a region that has a nearly uniform intensity. Figure 13 shows a feature extraction example using LBP.

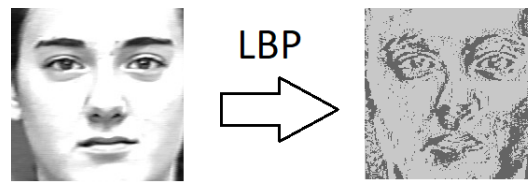


Figure 13. Feature extraction using LBP on a face from the CK+ database.

5.2. Optical Flow

OF [80] is a method that can only be used in a sequence of frames (video) since it aims to assess the magnitude and direction of motion. Basically, this technique calculates the motion between two frames at the pixel level. Therefore, it outputs a vector containing the movement of pixels in an image from the first frame to the second. However, this method depends on how well the initial tracked features are chosen and on how well they evolve through time, being highly sensitive to noise and occlusions [81].

This method can be effectively implemented in an FER system since, whenever one goes from a neutral expression to a peak facial expression, there is an obvious motion in the face that can be estimated. Figure 14 illustrates the usage of this method in an FER system using a sequence from the CK+ database.

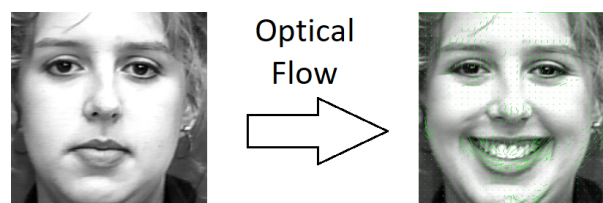


Figure 14. OF of a sequence from the CK+ database.

5.3. Active Appearance Model

Active appearance model (AAM) [82] is a Computer Vision algorithm that matches the object shape and appearance to a new image. In an FER system, this means matching only the face to a new image, everything else besides the face is discarded. The shape information extracted with the AAM is used to compute a set of suitable parameters that highlights the appearance of the facial features. However, this method is highly sensitive to input images that have differences in pose, expression, and illumination, which were not included in the training set [83]. Figure 15 shows an example of this process.



Figure 15. AAM shape estimation in a face from the CK+ database.

5.4. Action Units

AUs are individual muscle movements that constitute facial expressions. AUs were inspired by physiological, psychological, and sociological theories that claim that different facial expressions trigger different facial muscles. The Facial Action Coding System (FACS) [84] is a system to describe

facial expressions breaking them down in single AUs. Using AUs as features to classify emotions is a popular approach in the reviewed works, although some found difficulty in coding the dynamics of movements with precision, as well as measuring the AU intensity. Figure 16 illustrates a few examples of facial AUs.



Figure 16. Some relevant examples of AUs for facial expression discrimination.

5.5. Facial Animation Parameters

Facial Animation Parameters (FAPs) [85] represent 66 displacements and rotations of the feature points from the neutral face position. FAPs are based on facial motion and are related to muscle actions. They represent a complete set of basic facial actions, allowing the representation of facial expressions. FAPs can also be defined as relevant distances between facial features. However, FAPs extraction is sensitive to various noise sources, such as different lighting conditions, which might lead to subtle failures in facial area segmentation, compromising the FER system [86]. Figure 17 illustrates an example of FAPs extraction (eyebrows and mouth) using the CK+ database.

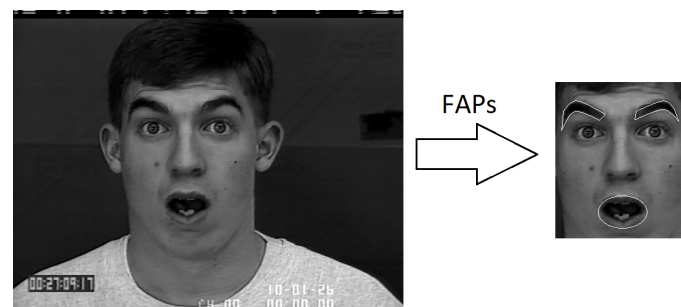


Figure 17. FAPs extraction using the CK+ database.

5.6. Gabor Filter

The Gabor filter [87] is used to represent texture information. It provides characteristic selection about orientation and scale, being robust to harsh illumination conditions. It is able to capture spatial information of frequency, position, and orientation from an image, and extract subtle local transformations effectively. However, the drawback of this method is the high-dimensional Gabor feature spaces, leading to a high computational cost, which is impractical for real-time applications. In order to have a real-time performance, one needs to use simplified Gabor features; however, these are sensitive to lighting variations [88]. Figure 18 illustrates a feature map of Gabor for a face from the CK+ database.

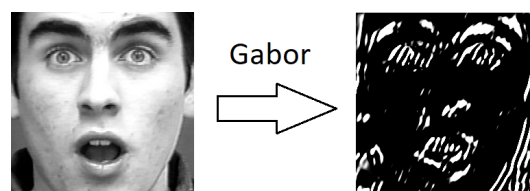


Figure 18. Feature map of Gabor for a face from the CK+ database.

5.7. Scale-Invariant Feature Transform

Scale-Invariant Feature Transform (SIFT) is a Computer Vision algorithm for detecting and describing local features in an image [89]. SIFT features are invariant to uniform scaling, orientation

and illumination changes; however, they suffer from blur and affine changes [90]. This robustness is because of the transformation of an image into a large collection of feature vectors, each of which is invariant to the conditions mentioned above. In FER systems, the SIFT algorithm can be used to detect facial features such as eyebrows, eyes, nose and mouth. Some works managed to combine this feature extraction algorithm with OF to calculate the motion of facial features. Figure 19 illustrates an example of extracting local features from a face of the CK+ database.

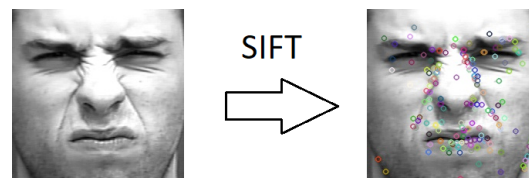


Figure 19. SIFT features of a face from the CK+ database.

5.8. Histogram of Oriented Gradients

Histogram of Oriented Gradients (HOG) [91] describes local object appearance and shape within an image by the distribution of intensity gradients or edge directions. The image is divided into cells with a determined number of pixels and a histogram of gradient directions is built for each cell. Since it operates on local cells, it is invariant to geometric transformations, except for object orientation. Figure 20 illustrates an example of HOG extraction using the CK+ database.

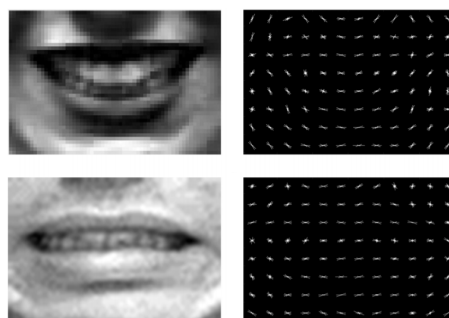


Figure 20. HOG descriptors of the mouth of two subjects from the CK+ database [92].

These HOG features are different and smoothly distinguished for each facial expression, which makes them appealing for FER systems.

6. Classification/Regression

A classification model is responsible for predicting a certain label given an input image or features. A regression model is responsible for determining the relationship between a dependent variable with independent variables. Both methods were used in the reviewed works, classification being the most popular one. The most used classification and regression algorithms are presented as follows.

6.1. Convolutional Neural Network

CNN is a type of neural networks that is mainly used in Computer Vision (Deep Learning) because of its ability to solve multiple image classification problems. CNNs can even beat humans in some of these problems since they are able to detect and identify underlying patterns that are too complex for the human eye. An input image is run through several hidden layers of the CNN that will decompose it into features. Those features are then used for classification, generally through a Softmax function that retrieves the highest probability from the classes' probability distribution as the predicted class.

It is important to mention that different problems require different CNN models and different tuning to maintain a high classification accuracy. This is mainly caused by the overfitting/underfitting

problem. Overfitting, as mentioned in Section 4.3.3, is when the model generalizes poorly; in other words, its classification accuracy drops considerably when classifying data that was not in the training set. Underfitting is when the model makes poor predictions in both training data and data it has not seen before. There are several ways to solve this problem:

1. Increasing the complexity of the model (by adding more layers).
2. Adding dropout layers [93] which randomly disable a determined number of nodes during training to avoid that the model memorizes patterns instead of learning them.
3. Tuning the parameters of the model during the training, such as epochs, batch size, learning rate, class weight, among others.
4. Increasing the training data by adding more samples or through DA as mentioned in Section 4.3.3.
5. Whenever the database is too small (a common problem on the publicly available databases for emotion recognition), Transfer Learning (TL) can be applied. TL uses a pre-defined model that has already been trained on a large database and one can fine-tune it using a smaller database for its own classification problem.

A great portion of the reviewed works used this approach for classification, showing promising results for Deep Learning based classifiers.

6.2. Support Vector Machine

SVM is a Machine Learning algorithm mostly used for classification/regression problems. An SVM model is the representation of features in space, mapped so that the features belonging to each class are divided by a clear gap that is as wide as possible. Input features are then mapped into that same space and predicted to belong to a class based on which side of the gap they fall. The training phase creates this map that is used after for predictions. The strengths of this classifier lie in handling complex nonlinear data and on being robust to overfitting. However, they are computationally expensive, hard to tune due to the importance of selecting the proper kernel function, and don't perform well with large databases. Figure 21 shows a potential map of an SVM model trained for an FER system.

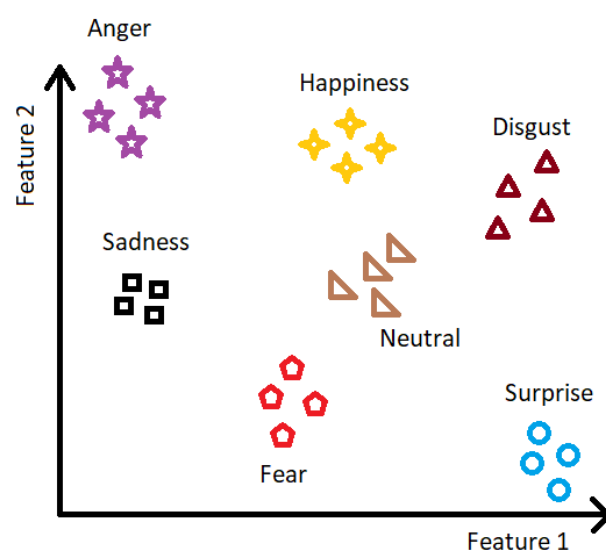


Figure 21. Example of a feature space in an FER system.

6.3. K-Nearest Neighbor

K-nearest neighbor (KNN) [94] is an instance-based learning algorithm that utilizes a non-parametric technique when making its classification or regression. The training data consist of vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists

only of storing these feature vectors and their belonging classes. In the classification step, an input feature or set of features are predicted by assigning the class that has the nearest features to the input. Common distance metrics used for calculating which features are closer to the input are the Euclidean distance (ED) and Hamming distance. The strengths of this classifier lie on the simple implementation and on the fast training step. However, it requires large storage space, the testing is slow, is sensitive to noise, and performs poorly for high-dimensional data. One more problem of this classification/regression approach is that unbalanced classes can lead to inaccurate predictions (classes with more number of samples usually dominate the predictions, even if wrongly). One way to overcome this problem is by setting class weights.

6.4. Naive Bayes

Naive Bayes classifiers [95] are a family of probabilistic Machine Learning classifiers based on Bayes theorem, which assume strong independence between features. The Bayes theorem is defined through the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \quad (1)$$

Using this equation, it is possible to calculate the likelihood of event A occurring given that B is true. However, these classifiers assume that features are independent, which means that, in an FER system, the Naive Bayes classifier will not correlate features when making a prediction. This is something undesirable since there are obvious correlated features when making facial expressions—for instance, when one is surprised, there is an obvious correlation between the mouth and the eyes (both usually open wider). However, the advantages of this classifier lie in the simple implementation and on scaling well for large databases. These type of classifiers are very popular in text classification problems.

6.5. Hidden Markov Model

Hidden Markov Model (HMM) [96] is a probabilistic model that is able to predict a sequence of unknown variables from a set of observed variables. For instance, in an FER system that would mean predicting happiness (hidden variable) based on a smile (observed variable). The strengths of this classifier lie on the potential to model arbitrary features from the observations, on the potential to merge various HMMs to classify more data and on the incorporation of previous knowledge into the model. However, this classifier is computationally expensive and struggles with overfitting.

6.6. Decision Tree

Decision Tree (DT) [97] as a classifier is basically a flowchart represented as a tree model. A DT splits the database into smaller sets of data until no more splits can be made, and the resulting leaves are the classes used for classification. The strengths of this classifier lie on the potential to learn nonlinear relationships of data, on handling high-dimensional data, and on the simple implementation. However, the main disadvantage of this classifier is overfitting, since it can keep branching until it memorizes the data during the training step.

6.7. Random Forest

Random Forest (RF) [98] is essentially an ensemble classifier, consisting in a group of DTs. Each DT outputs a prediction, and the final prediction is based on majority voting, meaning that the most predicted class will be the last prediction. It has the advantage of reducing overfitting over just one DT, since it reduces the bias by averaging the predictions of the ensemble. However, it has the disadvantage of becoming slower when increasing its complexity (e.g., by adding more DTs to the ensemble).

6.8. Euclidean Distance

ED is basically the distance between two points in Euclidean space. Some reviewed works used this metric for classification: calculating the distance from facial features of a certain facial expression to the mean vector of facial features for each emotion. The emotion that presents the closest distance is then assigned to the input face. The ED between two points (x,y) is defined through the following equation:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}. \quad (2)$$

The advantage of this classifier lies on its simple implementation to detect latent clusters. However, this simplicity is also its main drawback, especially for high-dimensional data.

7. Results and Discussion

In this section, some relevant results from the reviewed papers are presented. This aims to provide insights on what has been done in the FER field based on the reviewed works. However, it is difficult to make a fair comparison between works since most of them used different databases to train and test their FER systems, as well as different ratios of training/test data and different procedures. However, works that followed similar procedures and used the same databases can be compared. There are a few things to be aware of when analyzing the results. There are mainly two testing procedures used in the reviewed works:

- **Hold-out (HO)** is when the database is split up into two groups: one for training and one for testing. Generally, the training set has more data than the testing set (e.g., 70%/30%). This method has the advantage of being the fastest to be computed; however, it may produce high variance evaluations, since it heavily depends on which data end up being used for training and for testing.
- **Cross-validation (CV)**, which can be divided into:
 1. **K-fold cross-validation** [99], which is when the database is randomly split up into k groups. One group is for testing and the remaining for training. The process is repeated until every group is used for testing. This method has the advantage of being more robust to the division of the database, since every piece of data ends up being trained $k-1$ times and tested once. Therefore, the evaluation variance can be reduced as k is increased, although the computation time also increases.
 2. **Leave- p -out cross-validation** [100], which is when all possible sets of p data are left out from the training and used for validation. Although this method provides more robust evaluations than k -fold cross-validation, it can become computationally infeasible depending on p .

Within these two procedures, there is still the factor of being person-independent (PI): when the same person does not appear in the training set and in the testing set simultaneously. It is relevant to mention that testing procedures which followed a CV and PI approach are the most rigorous ones (meaning they usually have lower accuracy than the other testing procedures). The “Testing procedure” column, if it is not PI, means that the procedure potentially carried out with same people in the training set and testing set simultaneously. Another important remark is that some authors had common complaints about FER databases:

- The posed facial expressions made by actors when building the databases are too artificial. This means that, even if the works present a high accuracy on the benchmarks (using databases that are also built on posed facial expressions), it might not translate into a high accuracy when the same system faces a real world scenario.
- Some databases are poorly annotated or have an ambiguous annotation. Authors who noticed this problem tried to overcome it by making their own annotations or by excluding those samples.
- Emotion databases are generally small (mainly because of how hard it is to set up the image acquisition system and how hard it is to get several actors to do different facial expressions).

It is also worth noting that some classifiers, pre-processing, features, and databases presented in the tables were not approached in the previous sections. Since there is a huge variety of methods and databases used in the reviewed works, it was decided to only mention the most relevant ones. However, if the reader wants to know more about a specific work that used different techniques from the ones mentioned in the previous sections, it is recommended to check their work for a deeper understanding. The reviewed works are orderly referenced to facilitate potential searches. The results are compacted into tables and separated from their different approaches: static image (Table 2), video (Table 3), audiovisual (Table 4), video/static (Table 5) and circumplex model (Table 6). In the “Pre-processing” column, face detection is not present since every work did this ROI extraction technique. In Tables 2–5, the last column shows the accuracy obtained for each work in predicting emotions. Most works considered the six basic emotions, plus the neutral expression. However, not every work tried to make predictions in these six basic emotions, plus the neutral expression: some used five out of the six basic emotions, some added one more (contempt, usually seen on the CK+ database), and others went further and connected them (e.g., happily surprised). In Table 6, the last column shows the accuracy obtained for each work in predicting emotions using the circumplex model. This model has a wide variety of metrics, such as valence, arousal, power, dominance, expectation, and activation. Depending on the value of these metrics, one can predict which emotion certain subject is feeling. Figure 22 illustrates this concept using the valence and arousal metrics.

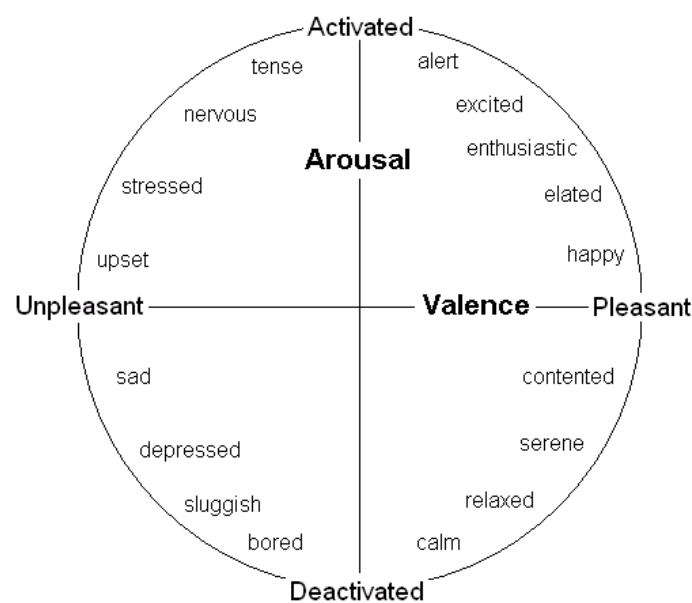


Figure 22. The 2D valence-arousal emotion space [101].

Tables 2–6 present the results of the reviewed works during this systematic review, sorted by accuracy. Some works tested their algorithms using various databases, but only the best classification result is presented in the tables.

Table 2 sums up results of works that approached this problem by analyzing static images. It is interesting to notice that the top results in this table mostly use CNNs with or without variants, CNNs combined with SVMs, and the top two did TL from pre-trained models. The works that attained the best results generally used a CNN pre-trained on large databases, being fine-tuned with smaller databases afterwards. This shows why CNNs have been dominating the FER field (as well as numerous Computer Vision fields): the joint optimization of both feature extraction and classification. Some used an end-to-end CNN approach, which is only fed with ROIs (faces) that could have been pre-processed or not. In other words, this approach does not need to choose which features to use for training their CNN: the CNN selects the features to learn by itself during the training step.

Table 2. Results of reviewed works for static image approaches.

Year	Classifier	Pre-processing	Features	Testing Procedure	Databases	Accuracy
[102] 2018	CNN (TL)	DA	ROI	CV (PI)	CK+/SFEW	98.90%
[103] 2017	WMDNN (TL)	Geometric/DA	LBP	CV	CK+/JAFPE/CASIA	97.02%
[104] 2018	DT/NN	Geometric	LBP	CV (PI)	CK+/JAFPE/...	96.48%
[105] 2019	CNN	Smooth	LBP/AUs	CV	CK+/JAFPE	96.46%
[106] 2019	CNN	DA	ROI	HO	FER-2013/JAFPE/...	96.44%
[107] 2018	CNN	Geometric	ROI	HO (PI)	CK+/RaFD/ADFES	96.27%
[108] 2018	CNN/SVM	-	Gabor	CV	KDEF	96.26%
[109] 2017	IACNN (TL)	Geometric/DRMF	ROI	CV (PI)	CK+/SFEW/MMI	95.37%
[110] 2016	RBF	HE	Curvelet	CV	CK/JAFPE	95.17%
[111] 2016	NN/SVM	HE/Smooth	LBP/Gabor	CV	CK+/MMI	94.66%
[112] 2017	SVM/CRF	-	AAM/Gabor	CV	CK+	93.93%
[113] 2018	CNN (TL)	Geometric/DA	Facial motions	CV (PI)	CK+/JAFPE/SFEW	93.64%
[114] 2018	RF	-	AUs/HOG	CV (PI)	CK+/BU4D/SFEW	93.40%
[115] 2016	KNN	-	Landmarks	HO	JAFPE/KDEF	92.29%
[116] 2008	ED	-	AAM	CV	FEEDTUM	91.70%
[117] 2018	SVM	CLAHE	DCT	HO	CK+/JAFPE/MMI	91.10%
[118] 2019	SVM	Smooth	CS-LOP	CV (PI)	CK+/JAFPE	89.58%
[119] 2019	CNN	Geometric/DA	ROI/AUs	HO	RaFD/AffectNet	81.83%
[120] 2017	CNN	Geometric/HE	ROI	HO	CK+/JAFPE	80.30%
[121] 2016	SVM	HDR	LBP/SURF	CV (PI)	JAFPE/SFEW	79.80%
[122] 2011	SVM	PCA/Geometric/Smooth	SIFT	CV (PI)	BU-3DFE	78.43%
[123] 2017	RF	-	Landmarks/HOG	HO	SFEW	75.39%
[124] 2019	DBN/SVM (TL)	-	ROI/OF	CV (PI)	BAUM/RML/MMI	73.73%
[125] 2010	HOSVD	-	Gabor/AUs	CV (PI)	CK/JAFPE	73.30%
[126] 2017	CNN	Geometric/DA	ROI	CV	FER-2013	71.03%
[127] 2015	CNN (TL)	Geometric/HE	ROI	HO	FER-2013/SFEW	61.29%
[128] 2015	CNN (TL)	Geometric	Landmarks	HO	FER-2013/SFEW	55.60%
[129] 2015	CNN (TL)	Geometric	LBP	HO	SFEW	54.56%
[130] 2016	SVM	-	AAM/AUs	CV	CK+	54.47%
[131] 2015	CNN (TL)	Geometric	LBP/HOG/...	CV	SFEW	51.08%

Other works used classifiers like DTs, RFs, or SVMs, instead of conventionally trying to overcome the problems of using CNN classifiers in FER (such as overfitting because of the overall small databases). These works showed that, with good understanding of the problem and of the used databases, as well as a proper pre-processing step (such as DA and intensity/spatial normalization), feature selection (such as AUs, LBP, HOG or Gabor features) and fine-tuning, it is possible to achieve competitive results.

However, there is a common discrepancy of accuracy when testing in controlled environment databases and in wild environment databases. This shows a clear difficulty in translating the good results in controlled environments (such as CK+ and JAFPE) to uncontrolled environments (such as FER-2013 and SFEW). Every work that tested its algorithms with various databases obtained a significantly worse result on the uncontrolled environment ones. One example from this table is the work [102], which despite obtaining 98.90% accuracy when testing on the CK+ database, only obtained 55.27% accuracy on the SFEW database. This is mainly caused by the head pose variation and the different lighting conditions to which a real world scenario is susceptible.

Table 3 sums up results of works that approached this problem by analyzing videos. It is possible to see the introduction of temporal features like OF and Motion History Histograms (MHHs). The pre-processing step revolves around intensity/spatial normalization and PCA. As for the classification step, there is a wide variety of classifiers that revolves around EDs, Gaussian Mixture Model (GMM), CNNs, SVMs and HMMs.

Table 3. Results of reviewed works for video approaches.

Year	Classifier	Pre-processing	Features	Testing Procedure	Databases	Accuracy
[132] 2016	ED	Geometric	Landmarks	HO	BU-4DFE/BP4D-S	100.00%
[133] 2017	GMM	Bandlet	LBP/KW	HO	CK/JAFFE	99.80%
[134] 2018	CNN	HE/Geometric	OF	HO	CK+/SAVEE/AFEW	98.77%
[135] 2018	DFSN-I	Geometric	AUs	CV (PI)	CK+/MMI/CASIA	98.73%
[136] 2013	SVM	-	AAM/AUs	CV	CK+	96.80%
[137] 2017	DBN	PCA	LDPP	CV	Depth	96.67%
[138] 2017	CNN	PCA	LDSP/LDRHP	CV	CK/Bosphorus	96.25%
[139] 2016	ED	-	Landmarks	CV	BU-4DFE	96.04%
[140] 2018	SVM	-	LBP	CV	CK+/MUG	95.80%
[141] 2017	TFP	LL	LBP	HO	CK/JAFFE	94.84%
[142] 2017	HMM	-	LBP/POMF	HO	Depth	94.17%
[143] 2012	SVM/RBF	-	OF/HOG	CV (PI)	CK	87.44%
[144] 2007	SVM	-	AAM/AUs	CV	CK	85.00%
[145] 2013	HMM	Geometric	Landmarks	CV (PI)	BU-4DFE	79.40%
[146] 2006	RNN	-	FAPs	HO	SAL	79.00%
[147] 2011	SVM	-	LBP/MHH	HO	GEMEP	70.30%
[148] 2007	Bayes	Intensity adjustment	Landmarks	CV (PI)	CK	70.20%
[149] 2015	SVM	-	LBP_TOP	CV (PI)	CASME II	69.63%
[150] 2019	CNN/LSTM (TL)	Geometric	ROI/OF	CV (PI)	RML/eNTERFACE'05	65.72%
[151] 2016	CNN/SVM	PCA/Geometric	LBP_TOP/SIFT	HO	AFEW	40.13%
[152] 2013	SVM	PCA	LBP/Gabor	HO	AFEW	30.05%
[153] 2013	SVM	Geometric	Gabor/AUs	HO	AFEW	29.81%

Most works tried to analyze facial expressions in a dynamic manner, by using the positional information of the detected facial features. In order to improve results in the video approach, the reviewed works tended to process spatial and temporal information separately. The spatial information can characterize different positions of facial features, while temporal information can capture the flow of these facial features through time (normally from a neutral expression up until a peak facial expression). Afterwards, they aggregated these two types of features and fed them into a single classifier or into a hybrid classifier. Normally, these classifiers need to have the capacity of retaining temporal information, for example by using Long Short-Term Memory Neural Networks (LSTMs) or Recurrent Neural Networks (RNNs).

The main conclusion to retrieve from video approaches is that retaining the temporal information usually leads to better results than analyzing each frame separately. However, as in static image approaches, the works with the lowest accuracy rates are from pose-variant databases, which means that retaining the temporal information is still not enough to perform well in uncontrolled environments.

Table 4 sums up results of works that approached this problem by analyzing both the video and its audio individually, combining them for the final classification. The columns “Classifiers”, “Pre-processing”, and “Features” are solely representing the visual analysis. However, the “Accuracy” column presents the last result of their multimodal system: fusion between video and audio. This approach has similar remarks to the video approach. The main differences are how the reviewed works processed the audio modality and how they combined the audiovisual features for the final model. The fusion methods that were able to attain the best results were the Probability-Based Product Rule and the Bayes sum rule.

Table 4. Results of reviewed works for audiovisual approaches.

Year	Classifiers	Pre-Processing	Features	Testing Procedure	Databases	Accuracy
[154] 2015	Bayes	Geometric	MK	CV	eNTERFACE'05	98.00%
[155] 2013	SVM	Smooth	Gabor/PCA	CV	eNTERFACE'05	80.27%
[156] 2014	MLP/RBF	-	ITMI/QIM	HO (PI)	CK/eNTERFACE'05	77.78%
[157] 2010	NN	-	FAPs/OF	HO (PI)	eNTERFACE'05	75.00%
[158] 2018	ED/CNN/LSTM (TL)	Geometric/DA	Landmarks	CV	AFEW/STED	61.87%
[159] 2008	SVM	PCA	Landmarks	HO	eNTERFACE'05	57.00%
[160] 2016	CNN (TL)	-	ROI	HO	FER-2013/AFEW	53.90%
[161] 2015	SVM	DCT/Geometric	AUs	HO	EmotiW	53.80%
[162] 2015	CNN	PCA	SIFT/LBP/...	CV	AFEW	53.62%
[163] 2015	CNN/RNN (TL)	HE/DA	ROI	HO	TFD/FER-2013/AFEW	52.88%
[164] 2017	CNN	Geometric	ROI	CV (PI)	AFEW/FER-2013/...	49.92%
[165] 2016	SVM/RF	-	LBP_TOP/AUs	CV	AFEW	46.88%
[166] 2014	SVM	Geometric	HOG_TOP	HO	AFEW	45.21%
[167] 2009	SVM/NN	-	FAPs/OF	HO	eNTERFACE'05	45.00%
[168] 2014	SVM	Geometric	LBP_TOP	HO	AFEW	41.77%
[169] 2013	SVM/CNN (TL)	Smooth/Contrast	ROI	HO	TFD/AFEW	41.03%
[170] 2013	SVM/HMM	-	Gabor/OF	HO	AFEW	20.51%

As in static image and video approaches, works with the lowest accuracy rates mostly resulted from uncontrolled environment databases (FER-2013 and AFEW). However, their results concluded that a multimodal approach is generally better than a unimodal approach.

Table 5 sums up results of a work that approached this problem by analyzing both the video and their frames individually, combining them for the final classification. The main remarks made by the authors is that this combined approach leads to better results than analyzing both parts alone.

Table 5. Results of a reviewed work with a video/static multimodal approach.

Year	Classifier	Pre-Processing	Features	Testing Procedure	Databases	Accuracy
[171] 2010	SVM	LDA	Landmarks	CV	JAFPE	87.50%

Table 6 sums up results of works that approached this problem by analyzing the circumplex model. What is presented in the “Accuracy” column is an average of the calculated metrics in this model. One thing that one can immediately notice is that every work used a video approach. The main features used for this approach revolve around facial landmarks and LBP features. As for the classification step, the main classifiers revolve around LSTMs, SVMs, HMMs, KNNs, and Support Vector Regressions (SVR).

Table 6. Results of reviewed works based on the circumplex model.

Year	Classifier	Approach	Pre-Processing	Features	Testing Procedure	Databases	Accuracy
[172] 2011	LSTM/HMM/SVM	Video	-	Landmarks	CV (PI)	SAL	85.00%
[173] 2014	NN	Audiovisual	Geometric/PCA	Landmarks	HO	AVEC2013	54.99%
[174] 2013	KNN/HMM	Audiovisual	PCA	LBP/AAM	CV (PI)	AVEC2011/...	52.60%
[175] 2013	SVR	Audiovisual	-	HOG/Harris3D	HO	AVEC2012	41.70%
[176] 2015	KNN/SVR	Video	-	LBP/EOH/LPQ	CV	AVEC2012/2013	14.09%

As in the video approach, in order to improve the results, the reviewed works tried to track facial features over time, processing their spatial and temporal information separately. By using a hybrid classifier with the capacity of retaining temporal information (LSTM), with a proper pre-processing step (e.g., intensity normalization which was neglected in the reviewed works of this table), it is possible to achieve competitive results when determining the metrics of the circumplex model. Some

remarks made by the authors are that the regions close to the eyes, mouth and eyebrows have dominant influence on FER, and that people tend to avoid eye contact when they do not want to talk about some topics.

8. Insights on Emotion Recognition in the Wild Challenge

Since the main observed problems with FER systems from this systematic review are pose-variant faces and wild environments, this section explores a yearly FER challenge in the wild (EmotiW). Most FER systems' goal is to face real world scenarios; therefore, it is important to analyze this challenge and how the participants are trying to tackle these environment adversities to pinpoint future directions. The EmotiW Challenge is divided into two different competitions: one is based in a static approach, using the SFEW database, and the other one is based in an audiovisual approach, using the AFEW database. In Section 7, it is possible to observe the overall superiority of multimodal approaches against the unimodal approaches; therefore, this section explores the audiovisual competition of the EmotiW Challenge.

Kahou et al. [169] proposed the combination of multiple emotion classifiers each based on a different data source. They used a CNN for frame-based classification of facial expressions from aligned faces. To classify whole video clips, they have implemented a video frame aggregation strategy based on SVMs. They have employed a shallow network architecture that focuses on extracted features of the mouth of the primary human subject in the scene and use these features as input to a SVM. Finally, they presented a novel technique to aggregate models based on random hyper-parameter search using low complexity aggregation techniques consisting of simple weighted averages to combine the visual model with the audio model. As for the pre-processing step, they did intensity normalization and isotropic smoothing. This work won the EmotiW 2013 challenge with the best submission achieving 41.03% accuracy.

Liu et al. [177] proposed to represent the AFEW video clips using three kinds of image set models: linear subspace, covariance matrix, and Gaussian distribution, respectively. Then, different kernels were employed on these set models correspondingly for distance measurement. Three classifiers were used: SVM, logistic regression, and partial least squares. Finally, a score-level fusion of classifiers based on different kernel methods and different modalities was conducted to further improve the performance. As for the feature extraction step, they extracted HOG and SIFT features, and used a CNN to exploit the strong spatially local correlations presented in the faces. This work won the EmotiW 2014 challenge with the best submission achieving 50.40% accuracy.

Yao et al. [161] proposed a pair-wise learning strategy to automatically seek a set of facial image patches which are important for discriminating two particular emotion categories called AU-aware facial features. In each pair-wise task, they used an undirected graph structure, which takes learnt facial patches as individual vertices, to encode feature relations between any two learnt facial patches. Finally, they constructed the emotion representation by concatenating all facial feature relations. As for the pre-processing step, they used a face frontalization method to remove the influence of head pose variation by normalizing their faces geometrically, and implemented a Discrete Cosine Transform (DCT) based method to compensate for illumination variations in the logarithm domain. In the feature extraction step, they extracted AUs and used Supervised Descent Method (SDM) to track them. This work won the EmotiW 2015 challenge with the best submission achieving 53.80% accuracy.

Fan et al. [178] proposed a hybrid network that combines LSTM and 3D convolutional networks (C3D). LSTM takes appearance features extracted by the CNN over individual video frames as input and encodes motion later, while C3D models appearance and motion of video simultaneously. This work emphasized the importance of pre-processing the data by testing their system with the original video frames, obtaining an average accuracy of just 20%. Therefore, as for the pre-processing step, they did face alignment. This work won the EmotiW 2016 challenge with the best submission achieving 59.02% accuracy.

Hu et al. [179] proposed a new learning method named Supervised Scoring Ensemble (SSE) with deep CNNs. They added supervision not only to deep layers but also to intermediate layers and

shallow layers to ease the training. They also presented a new fusion structure in which class-wise scoring activation at diverse complementary feature layers are concatenated and further used as the inputs for second-level supervision, acting as a deep feature ensemble within a single CNN architecture. As for the pre-processing step, they applied SDM to track facial features, face frontalization, rescaling, and applied a DCT based method for intensity normalization. As for the feature extraction step, they combined the grayscale face image with its corresponding basic LBP and mean LBP feature maps to form a three-channel input. This work won the EmotiW 2017 challenge with the best submission achieving 60.34% accuracy.

Liu et al. [158] proposed a hybrid net containing three main parts for the visual features: Landmark ED, CNN, and LSTM. These parts were then combined with weights for the final classification. For the Landmark ED part, 34 EDs were calculated as well as the mean, maximum, and variance, resulting in 102 total features for each video. This part alone achieved 39.95% accuracy on the AFEW database. For the CNN part, four CNNs were fine-tuned to predict single static images. Features extracted from these four CNNs were used to train a linear SVM. The VGG-Face model [180] was fine-tuned using the FER-2013 database for feature extraction. Those features were then used to train an LSTM, and this part alone was able to achieve 46.21% accuracy. Finally, they combined these three visual parts with the audio part as a final decision step. As for the pre-processing step, DA and face alignment were performed. This work won the EmotiW 2018 challenge with the best submission achieving 61.87% accuracy.

Figure 23 illustrates the EmotiW Challenge winners' accuracy over time.

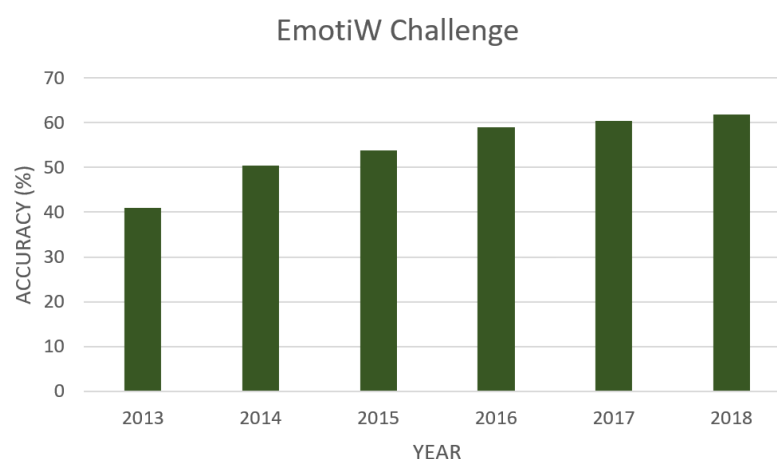


Figure 23. EmotiW challenge winners' accuracy over time.

It is possible to see that the EmotiW Challenge is stimulating the research done in emotion recognition in the wild, correlating with better results over the years. Recent winners used face frontalization to overcome the pose-variant faces problem. Since the AFEW database is 2D, it is the most reasonable solution. Overall, the winners don't seem to overlook the pre-processing step: most do geometric transformations and illumination corrections in order to normalize the data. Concerning the feature extraction step, at least two winners applied SDM to track facial features. The main facial features that were explored were AUs, HOG, SIFT, and LBP features. As for the classification step, they mainly used SVMs and fine-tuned CNNs in a combined way. Some winners also used LSTMs to make predictions based on the temporal data of the AFEW database.

Based on the EmotiW Challenge results, the future direction for FER in uncontrolled environments seems to be converging into:

- Pre-processing techniques that normalize pose-variant faces as well as the image intensity.
- The exploration of AUs, HOG, SIFT, and LBP features.
- The use of hybrid classifiers based on SVMs, fine-tuned CNNs, and LSTMs.

9. Conclusions

The interest in FER is growing gradually, and, with it, new algorithms and approaches are being developed. The recent popularization of Machine Learning made an obvious breakthrough in the research field. The research in FER is definitely in the right path, walking together with important fields like psychology, sociology, and physiology. From this, more and more accurate FER systems are emerging every year. However, despite this obvious progress, pose-variant faces in the wild are still a big challenge for FER systems. However, there are emotion recognition challenges every year that explores this problem and, with it, FER systems are becoming robust to pose-variant scenarios. Especially after a major breakthrough done by a CNN called AlexNet [181], which achieved a top-5 error of 15.31% in the ImageNet 2012 competition, more than 10.8% points lower than that of the runner up. After this, researchers became aware of the potential in CNNs for solving Computer Vision problems, and more FER systems using CNNs emerged, correlating with overall better results. The only potential negative aspect to point out from the reviewed works is that none considered the environment context. Although most works are giving the right steps towards multimodal systems, the environment context seems to be ignored. For instance, if there is an image of a birthday party, the happy context has a huge weight in the mood of people participating in it, which can't be ignored even if a certain participant is not explicitly smiling. Nevertheless, FER systems are being stimulated by yearly challenges and by the overall interest in numerous fields, achieving better results year by year.

Author Contributions: Conceptualization, D.C. and A.J.R.N.; Methodology, D.C.; Investigation, D.C.; Data Curation, D.C.; Writing—Original Draft Preparation, D.C.; Writing—Review and Editing, D.C. and A.J.R.N.; Visualization, D.C.; Supervision, A.J.R.N.; Project Administration, A.J.R.N.; Funding Acquisition, Integrated Programme of SR&TD SOCA (Ref. CENTRO-01-0145-FEDER-000010).

Funding: This research was funded by the Integrated Programme of SR&TD SOCA (Ref. CENTRO-01-0145-FEDER-000010), co-funded by the Centro 2020 program, Portugal 2020, European Union, through the European Regional Development Fund.

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

References

1. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
2. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436, doi:10.1038/nature14539.
3. Coan, J.A.; Allen, J.J. Frontal EEG asymmetry as a moderator and mediator of emotion. *Biol. Psychol.* **2004**, *67*, 7–50, doi:10.1016/j.biopsycho.2004.03.002.
4. Zafeiriou, S.; Zhang, C.; Zhang, Z. A survey on face detection in the wild: Past, present and future. *Comput. Vis. Image Underst.* **2015**, *138*, 1–24, doi:10.1016/j.cviu.2015.03.015.
5. Perez, L.; Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv* **2017**, arXiv:1712.04621.
6. Tian, Y.I.; Kanade, T.; Cohn, J.F. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 97–115, doi:10.1109/34.908962.
7. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, 27–29 July 1992; ACM: New York, NY, USA, 1992; pp. 144–152, doi:10.1145/130385.130401.
8. Liu, W.; Wang, Z.; Liu, X.; Zeng, N.; Liu, Y.; Alsaadi, F.E. A survey of deep neural network architectures and their applications. *Neurocomputing* **2017**, *234*, 11–26, doi:10.1016/j.neucom.2016.12.038.
9. ACM Digital Library. Available online: <https://dl.acm.org/> (accessed on 26 September 2019).
10. IEEE Xplore Digital Library. Available online: <https://ieeexplore.ieee.org/Xplore/home.jsp> (accessed on 26 September 2019).
11. Bielefeld Academic Search Engine. Available online: <https://www.base-search.net/> (accessed on 26 September 2019).
12. Springer Link. Available online: <https://link.springer.com/> (accessed on 26 September 2019).

13. Valstar, M.F.; Pantic, M.; Ambadar, Z.; Cohn, J.F. Spontaneous vs. posed facial behavior: Automatic analysis of brow actions. In Proceedings of the 8th International Conference on Multimodal Interfaces, Banff, AB, Canada, 2–4 November 2006; ACM: New York, NY, USA, 2006; pp. 162–170, doi:10.1145/1180995.1181031.
14. Duthoit, C.J.; Sztynka, T.; Lal, S.K.; Jap, B.T.; Agbinya, J.I. Optical flow image analysis of facial expressions of human emotion: Forensic applications. In Proceedings of the 1st International Conference on Forensic Applications and Techniques in Telecommunications, Information, and Multimedia and Workshop, Adelaide, Australia, 21–23 January 2008; p. 5.
15. Dornaika, F.; Davoine, F. Simultaneous facial action tracking and expression recognition in the presence of head motion. *Int. J. Comput. Vis.* **2008**, *76*, 257–281, doi:10.1007/s11263-007-0059-7.
16. Caridakis, G.; Karpouzis, K.; Kollias, S. User and context adaptive neural networks for emotion recognition. *Neurocomputing* **2008**, *71*, 2553–2562, doi:10.1016/j.neucom.2007.11.043.
17. Sun, X.; Rothkrantz, L.; Datcu, D.; Wiggers, P. A Bayesian approach to recognise facial expressions using vector flows. In Proceedings of the International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing, Ruse, Bulgaria, 18–19 June 2009; ACM: New York, NY, USA, 2009; p. 28, doi:10.1145/1731740.1731772.
18. Popa, M.; Rothkrantz, L.; Wiggers, P. Products appreciation by facial expressions analysis. In Proceedings of the 11th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing on International Conference on Computer Systems and Technologies, Sofia, Bulgaria, 17–18 June 2010; ACM: New York, NY, USA, 2010; pp. 293–298, doi:10.1145/1839379.1839431.
19. Liu, X.; Zhang, L.; Yadegar, J. A multi-modal emotion recognition system for persistent and non-invasive personal health monitoring. In Proceedings of the 2nd Conference on Wireless Health, La Jolla, 10–13 October 2011; ACM: New York, NY, USA, 2011; p. 28, doi:10.1145/2077546.2077577.
20. Metri, P.; Ghorpade, J.; Butalia, A. Facial emotion recognition using context based multimodal approach. *Int. J. Emerg. Sci.* **2012**, *2*, 171–183.
21. Cruz, A.C.; Bhanu, B.; Thakoor, N. Facial emotion recognition with expression energy. In Proceedings of the 14th ACM International Conference on Multimodal Interaction, Santa Monica, CA, USA, 22–26 October 2012; ACM: New York, NY, USA, 2012; pp. 457–464, doi:10.1145/2388676.2388777.
22. Soladié, C.; Salam, H.; Pelachaud, C.; Stoiber, N.; Séguier, R. A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection. In Proceedings of the 14th ACM International Conference on Multimodal Interaction, Santa Monica, CA, USA, 22–26 October 2012; ACM: New York, NY, USA, 2012; pp. 493–500, doi:10.1145/2388676.2388782.
23. Monkaresi, H.; Calvo, R.A.; Hussain, M.S. Automatic natural expression recognition using head movement and skin color features. In Proceedings of the International Working Conference on Advanced Visual Interfaces, Capri Island, Italy, 21–25 May 2012; ACM: New York, NY, USA, 2012; pp. 657–660, doi:10.1145/2254556.2254678.
24. Biel, J.I.; Teijeiro-Mosquera, L.; Gatica-Perez, D. Facetube: Predicting personality from facial expressions of emotion in online conversational video. In Proceedings of the 14th ACM International Conference on Multimodal Interaction, Santa Monica, CA, USA, 22–26 October 2012; ACM: New York, NY, USA, 2012; pp. 53–56, doi:10.1145/2388676.2388689.
25. Nedkov, S.; Dimov, D. Emotion recognition by face dynamics. In Proceedings of the 14th International Conference on Computer Systems and Technologies, Ruse, Bulgaria, 28–29 June 2013; ACM: New York, NY, USA, 2013; pp. 128–136, doi:10.1145/2516775.2516794.
26. Terzis, V.; Moridis, C.N.; Economides, A.A. Measuring instant emotions based on facial expressions during computer-based assessment. *Pers. Ubiquitous Comput.* **2013**, *17*, 43–52, doi:10.1007/s00779-011-0477-y.
27. Meng, H.; Huang, D.; Wang, H.; Yang, H.; Ai-Shuraifi, M.; Wang, Y. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge, Barcelona, Spain, 21–25 October 2013; ACM: New York, NY, USA, 2013; pp. 21–30, doi:10.1145/2512530.2512532.
28. Gómez Jáuregui, D.A.; Martín, J.C. Evaluation of vision-based real-time measures for emotions discrimination under uncontrolled conditions. In Proceedings of the 2013 on Emotion Recognition in the Wild Challenge and Workshop, Sydney, Australia, 9–13 December 2013; ACM: New York, NY, USA, 2013; pp. 17–22, doi:10.1145/2531923.2531925.

29. Bakhtiyari, K.; Husain, H. Fuzzy model of dominance emotions in affective computing. *Neural Comput. Appl.* **2014**, *25*, 1467–1477, doi:10.1007/s00521-014-1637-6.
30. Sangineto, E.; Zen, G.; Ricci, E.; Sebe, N. We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; ACM: New York, NY, USA, 2014; pp. 357–366, doi:10.1145/2647868.2654916.
31. Jang, G.J.; Jo, A.; Park, J.S. Video-based emotion identification using face alignment and support vector machines. In Proceedings of the Second International Conference on Human-agent Interaction, Tsukuba, Japa, 28–31 October 2014; ACM: New York, NY, USA, 2014; pp. 285–286, doi:10.1145/2658861.2658943.
32. Zen, G.; Sangineto, E.; Ricci, E.; Sebe, N. Unsupervised domain adaptation for personalized facial emotion recognition. In Proceedings of the 16th International Conference on Multimodal Interaction, Istanbul, Turkey, 12–16 November 2014; ACM: New York, NY, USA, 2014; pp. 128–135, doi:10.1145/2663204.2663247.
33. Rothkrantz, L. Online emotional facial expression dictionary. In Proceedings of the 15th International Conference on Computer Systems and Technologies, Ruse, Bulgaria, 27–28 June 2014; ACM: New York, NY, USA, 2014; pp. 116–123, doi:10.1145/2659532.2659627.
34. Chao, L.; Tao, J.; Yang, M.; Li, Y.; Wen, Z. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, Brisbane, Australia, 26–30 October 2015; ACM: New York, NY, USA, 2015; pp. 65–72, doi:10.1145/2808196.2811634.
35. Kim, Y.; Provost, E.M. Emotion recognition during speech using dynamics of multiple regions of the face. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2015**, *12*, 25, doi:10.1145/2808204.
36. Nomiya, H.; Sakaue, S.; Hochin, T. Recognition and intensity estimation of facial expression using ensemble classifiers. In Proceedings of the 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, Japan, 26–29 June 2016; pp. 1–6, doi:10.1109/ICIS.2016.7550861.
37. Zhang, Y.D.; Yang, Z.J.; Lu, H.M.; Zhou, X.X.; Phillips, P.; Liu, Q.M.; Wang, S.H. Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation. *IEEE Access* **2016**, *4*, 8375–8385, doi:10.1109/ACCESS.2016.2628407.
38. Barsoum, E.; Zhang, C.; Ferrer, C.C.; Zhang, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; ACM: New York, NY, USA, 2016; pp. 279–283, doi:10.1145/2993148.2993165.
39. Liu, Z.; Wu, M.; Cao, W.; Chen, L.; Xu, J.; Zhang, R.; Zhou, M.; Mao, J. A facial expression emotion recognition based human-robot interaction system. *IEEE/CAA J. Autom. Sin.* **2017**, *4*, 668–676.
40. Bouzakraoui, M.S.; Sadiq, A.; Enneya, N. A Customer Emotion Recognition through Facial Expression using POEM descriptor and SVM classifier. In Proceedings of the 2nd International Conference on Big Data, Cloud and Applications, Tetouan, Morocco, 29–30 March 2017; ACM: New York, NY, USA, 2017; p. 80.
41. Elfaramawy, N.; Barros, P.; Parisi, G.I.; Wermter, S. Emotion recognition from body expressions with a neural network architecture. In Proceedings of the 5th International Conference on Human Agent Interaction, Bielefeld, Germany, 17–20 October 2017; ACM: New York, NY, USA, 2017; pp. 143–149, doi:10.1145/3125739.3125772.
42. Qi, C.; Li, M.; Wang, Q.; Zhang, H.; Xing, J.; Gao, Z.; Zhang, H. Facial expressions recognition based on cognition and mapped binary patterns. *IEEE Access* **2018**, *6*, 18795–18803, doi:10.1109/ACCESS.2018.2816044.
43. Zhang, Z.; Chen, T.; Meng, H.; Liu, G.; Fu, X. SMEConvnet: A convolutional neural network for spotting spontaneous facial micro-expression from long videos. *IEEE Access* **2018**, *6*, 71143–71151, doi:10.1109/ACCESS.2018.2879485.
44. Guo, J.; Lei, Z.; Wan, J.; Avots, E.; Hajarolasvadi, N.; Knyazev, B.; Kuharenko, A.; Junior, J.C.S.J.; Baró, X.; Demirel, H.; et al. Dominant and complementary emotion recognition from still images of faces. *IEEE Access* **2018**, *6*, 26391–26403, doi:10.1109/ACCESS.2018.2831927.
45. Slimani, K.; Kas, M.; El Merabet, Y.; Messoussi, R.; Ruichek, Y. Facial emotion recognition: A comparative analysis using 22 LBP variants. In Proceedings of the 2nd Mediterranean Conference on Pattern Recognition and Artificial Intelligence, Rabat, Morocco, 27–28 March 2018; ACM: New York, NY, USA, 2018; pp. 88–94, doi:10.1145/3177148.3180092.

46. Bernin, A.; M'uller, L.; Ghose, S.; Grecos, C.; Wang, Q.; Jettke, R.; von Luck, K.; Vogt, F. Automatic Classification and Shift Detection of Facial Expressions in Event-Aware Smart Environments. In Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference, Corfu, Greece, 26–29 June 2018; ACM: New York, NY, USA, 2018; pp. 194–201, doi:10.1145/3197768.3201527.
47. Magdin, M.; Prikler, F. Real time facial expression recognition using webcam and SDK affectiva. *IJIMAI* **2018**, *5*, 7–15.
48. Pham, T.T.D.; Kim, S.; Lu, Y.; Jung, S.W.; Won, C.S. Facial action units-based image retrieval for facial expression recognition. *IEEE Access* **2019**, *7*, 5200–5207, doi:10.1109/ACCESS.2018.2889852.
49. Slimani, K.; Lekdioui, K.; Messoussi, R.; Touahni, R. Compound Facial Expression Recognition Based on Highway CNN. In Proceedings of the New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society, Kenitra, Morocco, 28–29 March 2019; ACM: New York, NY, USA, 2019; p. 1, doi:10.1145/3314074.3314075.
50. 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, doi:10.1109/CVPRW.2010.5543262.
51. Lyons, M.J.; Akamatsu, S.; Kamachi, M.; Gyoba, J.; Budynek, J. The Japanese female facial expression (JAFPE) database. In Proceedings of the Third International Conference on Automatic Face And Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 14–16, doi:10.5281/zenodo.3451524.
52. Yin, L.; Wei, X.; Sun, Y.; Wang, J.; Rosato, M.J. A 3D facial expression database for facial behavior research. In Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06), Southampton, UK, 10–12 April 2006; pp. 211–216, doi:10.1109/FGR.2006.6.
53. 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, Germany, 2013; pp. 117–124, doi:10.1007/978-3-642-42051-1_16.
54. Dhall, A.; Ramana Murthy, O.; Goecke, R.; Joshi, J.; Gedeon, T. Video and image based emotion recognition challenges in the wild: EmotiW 2015. In Proceedings of the 2015 ACM on International Conference On Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; ACM: New York, NY, USA, 2015; pp. 423–426, doi:10.1145/2818346.2829994.
55. 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, Netherlands, 6 July 2005; pp. 5–pp, doi:10.1109/ICME.2005.1521424.
56. Martin, O.; Kotsia, I.; Macq, B.; Pitas, I. The eINTERFACE'05 audio-visual emotion database. In Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA, 3–7 April 2006; p. 8, doi:10.1109/ICDEW.2006.145.
57. Calvo, M.G.; Lundqvist, D. Facial expressions of emotion (KDEF): Identification under different display-duration conditions. *Behav. Res. Methods* **2008**, *40*, 109–115, doi:10.3758/BRM.40.1.109.
58. Langner, O.; Dotsch, R.; Bijlstra, G.; Wigboldus, D.H.; Hawk, S.T.; Van Knippenberg, A. Presentation and validation of the Radboud Faces Database. *Cogn. Emot.* **2010**, *24*, 1377–1388, doi:10.1080/02699930903485076.
59. The Extended Cohn–Kanade Database. Available online: <http://www.consortium.ri.cmu.edu/ckagree/> (accessed on 8 September 2019).
60. The Japanese Female Facial Expression Database. Available online: <http://www.kasrl.org/jaffe.html> (accessed on 8 September 2019).
61. Binghamton University 3D Facial Expression Database. Available online: http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html (accessed on 8 September 2019).
62. Facial Expression Recognition 2013 Database. Available online: <https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data> (accessed on 8 September 2019).
63. Emotion Recognition in the Wild Database. Available online: <https://cs.anu.edu.au/few/AFEW.html> (accessed on 8 September 2019).
64. MMI Database. Available online: <https://mmifacedb.eu/> (accessed on 8 September 2019).
65. eINTERFACE'05 Audio-Visual Emotion Database. Available online: <http://www.interface.net/interface05/> (accessed on 8 September 2019).

66. Karolinska Directed Emotional Faces Database. Available online: <http://kdef.se/> (accessed on 8 September 2019).
67. Radboud Faces Database. Available online: <http://www.socsci.ru.nl:8180/RaFD2/RaFD?p=main> (accessed on 8 September 2019).
68. Viola, P.; Jones, M.J. Robust real-time face detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154, doi:10.1023/B:VISI.0000013087.49260.fb.
69. Kazemi, V.; Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; 2014; pp. 1867–1874.
70. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503, doi:10.1109/LSP.2016.2603342.
71. Farfadi, S.S.; Saberian, M.J.; Li, L.J. Multi-view face detection using deep convolutional neural networks. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, 23–26 June 2015; ACM: New York, NY, USA, 2015; pp. 643–650, doi:10.1145/2671188.2749408.
72. Azulay, A.; Weiss, Y. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv* **2018**, arXiv:1805.12177.
73. Tomasi, C.; Manduchi, R. Bilateral filtering for gray and color images. *ICCV* **1998**, *98*, 2, doi:10.1109/ICCV.1998.710815.
74. Lindenbaum, M.; Fischer, M.; Bruckstein, A. On Gabor's contribution to image enhancement. *Pattern Recognit.* **1994**, *27*, 1–8, doi:10.1016/0031-3203(94)90013-2.
75. Garg, P.; Jain, T. A Comparative Study on Histogram Equalization and Cumulative Histogram Equalization. *Int. J. New Technol. Res.* **2017**, *3*, 41–43.
76. Pizer, S.M.; Amburn, E.P.; Austin, J.D.; Cromartie, R.; Geselowitz, A.; Greer, T.; ter Haar Romeny, B.; Zimmerman, J.B.; Zuiderveld, K. Adaptive histogram equalization and its variations. *Comput. Vis. Graph. Image Process.* **1987**, *39*, 355–368, doi:10.1016/S0734-189X(87)80186-X.
77. Hawkins, D.M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12, doi:10.1021/ci0342472.
78. Jolliffe, I. *Principal Component Analysis*; Springer: Berlin, Germany 2011, doi:10.1007/978-3-642-04898-2.
79. Ojala, T.; Pietikäinen, M.; Mäenpää, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987, doi:10.1109/TPAMI.2002.1017623.
80. Horn, B.K.; Schunck, B.G. Determining optical flow. *Artif. Intell.* **1981**, *17*, 185–203, doi:10.1016/0004-3702(81)90024-2.
81. Barron, J.L.; Fleet, D.J.; Beauchemin, S.S.; Burkitt, T. Performance of optical flow techniques. In Proceedings of the 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Champaign, IL, USA, 15–18 June 1992; pp. 236–242, doi:10.1109/CVPR.1992.223269.
82. Cootes, T.F.; Edwards, G.J.; Taylor, C.J. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 681–685, doi:10.1109/34.927467.
83. Abdulameer, M.H.; Abdullah, S.; Huda, S.N.; Othman, Z.A. A modified active appearance model based on an adaptive artificial bee colony. *Sci. World J.* **2014**, *2014*, doi:10.1155/2014/879031.
84. Ekman, R. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*; Oxford University Press: New York, NY, USA, 1997.
85. Pakstas, A.; Forchheimer, R.; Pandzic, I.S. *MPEG-4 Facial Animation: The Standard, Implementation and Applications*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2002.
86. Chandrasiri, N.P.; Naemura, T.; Harashima, H. Real time facial expression recognition system with applications to facial animation in MPEG-4. *IEICE Trans. Inf. Syst.* **2001**, *84*, 1007–1017.
87. Jain, A.K.; Farrokhnia, F. Unsupervised texture segmentation using Gabor filters. *Pattern Recognit.* **1991**, *24*, 1167–1186, doi:10.1016/0031-3203(91)90143-S.
88. Choraś, R.S. *Image Processing and Communications Challenges 2*; Springer: Berlin, Germany, 2010; pp. 15–17, doi:10.1007/978-3-642-16295-4.
89. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110, doi:10.1023/B:VISI.0000029664.99615.94.
90. Wu, J.; Cui, Z.; Sheng, V.S.; Zhao, P.; Su, D.; Gong, S. A Comparative Study of SIFT and its Variants. *Meas. Sci. Rev.* **2013**, *13*, 122–131, doi:10.2478/msr-2013-0021.

91. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; doi:10.1109/CVPR.2005.177.
92. Liu, Y.; Li, Y.; Ma, X.; Song, R. Facial expression recognition with fusion features extracted from salient facial areas. *Sensors* **2017**, *17*, 712, doi:10.3390/s17040712.
93. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
94. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27, doi:10.1109/TIT.1967.1053964.
95. Friedman, N.; Geiger, D.; Goldszmidt, M. Bayesian network classifiers. *Mach. Learn.* **1997**, *29*, 131–163, doi:10.1023/A:1007465528199.
96. Eddy, S.R. Hidden markov models. *Curr. Opin. Struct. Biol.* **1996**, *6*, 361–365, doi:10.1016/S0959-440X(96)80056-X.
97. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106, doi:10.1007/BF00116251.
98. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32, doi:10.1023/A:1010933404324.
99. Chui, K.T.; Lytras, M.D. A Novel MOGA-SVM Multinomial Classification for Organ Inflammation Detection. *Appl. Sci.* **2019**, *9*, 2284, doi:10.3390/app9112284.
100. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79, doi:10.1214/09-SS054.
101. Trimmer, P.; Paul, E.; Mendl, M.; McNamara, J.; Houston, A. On the evolution and optimality of mood states. *Behav. Sci.* **2013**, *3*, 501–521, doi:10.3390/bs3030501.
102. 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, doi:10.1007/s11263-017-1055-1.
103. Yang, B.; Cao, J.; Ni, R.; Zhang, Y. Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. *IEEE Access* **2017**, *6*, 4630–4640, doi:10.1109/ACCESS.2017.2784096.
104. Gogić, I.; Manhart, M.; Pandžić, I.S.; Ahlberg, J. Fast facial expression recognition using local binary features and shallow neural networks. *Vis. Comput.* **2018**, *1*–16, doi:10.1007/s00371-018-1585-8.
105. Kim, J.H.; Kim, B.G.; Roy, P.P.; Jeong, D.M. Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure. *IEEE Access* **2019**, *7*, 41273–41285, doi:10.1109/ACCESS.2019.2907327.
106. Hua, W.; Dai, F.; Huang, L.; Xiong, J.; Gui, G. HERO: Human emotions recognition for realizing intelligent Internet of Things. *IEEE Access* **2019**, *7*, 24321–24332, doi:10.1109/ACCESS.2019.2900231.
107. Wu, B.F.; Lin, C.H. Adaptive feature mapping for customizing deep learning based facial expression recognition model. *IEEE Access* **2018**, *6*, 12451–12461, doi:10.1109/ACCESS.2018.2805861.
108. Ruiz-Garcia, A.; Elshaw, M.; Altahhan, A.; Palade, V. A hybrid deep learning neural approach for emotion recognition from facial expressions for socially assistive robots. *Neural Comput. Appl.* **2018**, *29*, 359–373, doi:10.1007/s00521-018-3358-8.
109. Meng, Z.; Liu, P.; Cai, J.; Han, S.; Tong, Y. Identity-aware convolutional neural network for facial 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. 558–565, doi:10.1109/FG.2017.140.
110. Uçar, A.; Demir, Y.; Güzelış, C. A new facial expression recognition based on curvelet transform and online sequential extreme learning machine initialized with spherical clustering. *Neural Comput. Appl.* **2016**, *27*, 131–142, doi:10.1007/s00521-014-1569-1.
111. Mistry, K.; Zhang, L.; Neoh, S.C.; Lim, C.P.; Fielding, B. A micro-GA embedded PSO feature selection approach to intelligent facial emotion recognition. *IEEE Trans. Cybern.* **2016**, *47*, 1496–1509, doi:10.1109/TCYB.2016.2549639.
112. Liliana, D.Y.; Basaruddin, C.; Widianto, M.R. Mix emotion recognition from facial expression using SVM-CRF sequence classifier. In Proceedings of the International Conference on Algorithms, Computing and Systems, Jeju Island, Korea, 10–13 August 2017; ACM: New York, NY, USA, 2017; pp. 27–31, doi:10.1145/3127942.3127958.
113. Ferreira, P.M.; Marques, F.; Cardoso, J.S.; Rebelo, A. Physiological Inspired Deep Neural Networks for Emotion Recognition. *IEEE Access* **2018**, *6*, 53930–53943, doi:10.1109/ACCESS.2018.2870063.

114. Dapogny, A.; Bailly, K.; Dubuisson, S. Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection. *Int. J. Comput. Vis.* **2018**, *126*, 255–271, doi:10.1007/s11263-017-1010-1.
115. Yaddaden, Y.; Bouzouane, A.; Adda, M.; Bouchard, B. A new approach of facial expression recognition for ambient assisted living. In Proceedings of the 9th ACM International Conference on Pervasive Technologies Related to Assistive Environments, Corfu, Island, Greece, 29 June –1 July 2016; ACM: New York, NY, USA, 2016; p. 14, doi:10.1145/2910674.2910703.
116. Ratliff, M.S.; Patterson, E. Emotion recognition using facial expressions with active appearance models. In Proceedings of the Third IASTED International Conference on Human Computer Interaction, Innsbruck, Austria, 17–19 March 2008.
117. Khan, S.A.; Hussain, S.; Xiaoming, S.; Yang, S. An effective framework for driver fatigue recognition based on intelligent facial expressions analysis. *IEEE Access* **2018**, *6*, 67459–67468, doi:10.1109/ACCESS.2018.2878601.
118. Hu, M.; Zheng, Y.; Yang, C.; Wang, X.; He, L.; Ren, F. Facial Expression Recognition Using Fusion Features Based on Center-Symmetric Local Octonary Pattern. *IEEE Access* **2019**, *7*, 29882–29890, doi:10.1109/ACCESS.2019.2899024.
119. Deng, J.; Pang, G.; Zhang, Z.; Pang, Z.; Yang, H.; Yang, G. cGAN Based Facial Expression Recognition for Human-Robot Interaction. *IEEE Access* **2019**, *7*, 9848–9859, doi:10.1109/ACCESS.2019.2891668.
120. Shan, K.; Guo, J.; You, W.; Lu, D.; Bie, R. Automatic facial expression recognition based on a deep convolutional-neural-network structure. In Proceedings of the 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), London, UK, 7–9 June 2017; pp. 123–128, doi:10.1109/SERA.2017.7965717.
121. Ige, E.O.; Debattista, K.; Chalmers, A. Towards hdr based facial expression recognition under complex lighting. In Proceedings of the 33rd Computer Graphics International, Heraklion, Greece, 28 June–1 July 2016; ACM: New York, NY, USA, 2016; pp. 49–52, doi:10.1145/2949035.2949048.
122. Berretti, S.; Amor, B.B.; Daoudi, M.; Del Bimbo, A. 3D facial expression recognition using SIFT descriptors of automatically detected keypoints. *Vis. Comput.* **2011**, *27*, 1021, doi:10.1007/s00371-011-0611-x.
123. Rassadin, A.; Gruzdev, A.; Savchenko, A. Group-level emotion recognition using transfer learning from face identification. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; ACM: New York, NY, USA, 2017; pp. 544–548, doi:10.1145/3136755.3143007.
124. Zhang, S.; Pan, X.; Cui, Y.; Zhao, X.; Liu, L. Learning Affective Video Features for Facial Expression Recognition via Hybrid Deep Learning. *IEEE Access* **2019**, *7*, 32297–32304, doi:10.1109/ACCESS.2019.2901521.
125. Tan, H.; Zhang, Y.; Cheri, H.; Zhao, Y.; Wang, W. Person-independent expression recognition based on person-similarity weighted expression feature. *J. Syst. Eng. Electron.* **2010**, *21*, 118–126, doi:10.3969/j.issn.1004-4132.2010.01.019.
126. Sang, D.V.; Cuong, L.T.B.; Van Thieu, V. Multi-task learning for smile detection, emotion recognition and gender classification. In Proceedings of the Eighth International Symposium on Information and Communication Technology, Nha Trang City, Vietnam, 7–8 December 2017; ACM: New York, NY, USA, 2017; pp. 340–347, doi:10.1145/3155133.3155207.
127. 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; ACM: New York, NY, USA, 2015; pp. 435–442, doi:10.1145/2818346.2830595.
128. 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; ACM: New York, NY, USA, 2015; pp. 443–449, doi:10.1145/2818346.2830593.
129. Levi, G.; Hassner, T. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In Proceedings of the 2015 ACM on International Conference On Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; ACM: New York, NY, USA, 2015; pp. 503–510, doi:10.1145/2818346.2830587.
130. Sert, M.; Aksoy, N. Recognizing facial expressions of emotion using action unit specific decision thresholds. In Proceedings of the 2nd Workshop on Advancements in Social Signal Processing for Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; ACM: New York, NY, USA, 2016; pp. 16–21, doi:10.1145/3005467.3005471.

131. Sun, B.; Li, L.; Zhou, G.; Wu, X.; He, J.; Yu, L.; Li, D.; Wei, Q. Combining multimodal features within a fusion network for emotion recognition in the wild. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; ACM: New York, NY, USA, 2015; pp. 497–502, doi:10.1145/2818346.2830586.
132. Danelakis, A.; Theoharis, T.; Pratikakis, I. A spatio-temporal wavelet-based descriptor for dynamic 3D facial expression retrieval and recognition. *Vis. Comput.* **2016**, *32*, 1001–1011, doi:10.1007/s00371-016-1243-y.
133. Hossain, M.S.; Muhammad, G. An emotion recognition system for mobile applications. *IEEE Access* **2017**, *5*, 2281–2287, doi:10.1109/ACCESS.2017.2672829.
134. Zhao, J.; Mao, X.; Zhang, J. Learning deep facial expression features from image and optical flow sequences using 3D CNN. *Vis. Comput.* **2018**, *34*, 1461–1475, doi:10.1007/s00371-018-1477-y.
135. Tang, Y.; Zhang, X.M.; Wang, H. Geometric-convolutional feature fusion based on learning propagation for facial expression recognition. *IEEE Access* **2018**, *6*, 42532–42540, doi:10.1109/ACCESS.2018.2858278.
136. Stankovic, I.; Karnjanadecha, M. Use of septum as reference point in a neurophysiologic approach to facial expression recognition. *Songklanakarin J. Sci. Technol.* **2013**, *35*, 461–468.
137. Uddin, M.Z.; Hassan, M.M.; Almogren, A.; Alamri, A.; Alrubaian, M.; Fortino, G. Facial expression recognition utilizing local direction-based robust features and deep belief network. *IEEE Access* **2017**, *5*, 4525–4536, doi:10.1109/ACCESS.2017.2676238.
138. Uddin, M.Z.; Khaksar, W.; Torresen, J. Facial expression recognition using salient features and convolutional neural network. *IEEE Access* **2017**, *5*, 26146–26161, doi:10.1109/ACCESS.2017.2777003.
139. Danelakis, A.; Theoharis, T.; Pratikakis, I. A robust spatio-temporal scheme for dynamic 3D facial expression retrieval. *Vis. Comput.* **2016**, *32*, 257–269, doi:10.1007/s00371-015-1142-7.
140. Agarwal, S.; Santra, B.; Mukherjee, D.P. Anubhav: Recognizing emotions through facial expression. *Vis. Comput.* **2018**, *34*, 177–191, doi:10.1007/s00371-016-1323-z.
141. Ding, Y.; Zhao, Q.; Li, B.; Yuan, X. Facial expression recognition from image sequence based on LBP and Taylor expansion. *IEEE Access* **2017**, *5*, 19409–19419, doi:10.1109/ACCESS.2017.2737821.
142. Kabir, M.H.; Salekin, M.S.; Uddin, M.Z.; Abdullah-Al-Wadud, M. Facial expression recognition from depth video with patterns of oriented motion flow. *IEEE Access* **2017**, *5*, 8880–8889, doi:10.1109/ACCESS.2017.2704087.
143. Agarwal, S.; Chatterjee, M.; Mukherjee, P.D. Recognizing facial expressions using a novel shape motion descriptor. In Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing, Mumbai, India, 16–19 December 2012; ACM: New York, NY, USA, 2012; p. 29, doi:10.1145/2425333.2425362.
144. Datcu, D.; Rothkrantz, L. Facial expression recognition in still pictures and videos using active appearance models: A comparison approach. In Proceedings of the 2007 International Conference on Computer Systems and Technologies, Bulgaria, 14–15 June 2007; ACM: New York, NY, USA, 2007; p. 112, doi:10.1145/1330598.1330717.
145. Berretti, S.; Del Bimbo, A.; Pala, P. Automatic facial expression recognition in real-time from dynamic sequences of 3D face scans. *Vis. Comput.* **2013**, *29*, 1333–1350, doi:10.1007/s00371-013-0869-2.
146. Caridakis, G.; Malatesta, L.; Kessous, L.; Amir, N.; Raouzaoui, A.; Karpouzis, K. Modeling naturalistic affective states via facial and vocal expressions recognition. In Proceedings of the 8th International Conference on Multimodal Interfaces, Banff, AB, Canada, 2–4 November 2006; ACM: New York, NY, USA, 2006; pp. 146–154, doi:10.1145/1180995.1181029.
147. Meng, H.; Romera-Paredes, B.; Bianchi-Berthouze, N. Emotion recognition by two view SVM_2K classifier on dynamic facial expression features. In Proceedings of the Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011), Santa Barbara, CA, USA, 21–25 March 2011; pp. 854–859, doi:10.1109/FG.2011.5771362.
148. Kumano, S.; Otsuka, K.; Yamato, J.; Maeda, E.; Sato, Y. Pose-invariant facial expression recognition using variable-intensity templates. In *Asian Conference on Computer Vision*; Springer: Berlin, Germany, 2007; pp. 324–334, doi:10.1007/978-3-540-76386-4_30.
149. Park, S.Y.; Lee, S.H.; Ro, Y.M. Subtle facial expression recognition using adaptive magnification of discriminative facial motion. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; ACM: New York, NY, USA, 2015; pp. 911–914, doi:10.1145/2733373.2806362.

150. Pan, X.; Ying, G.; Chen, G.; Li, H.; Li, W. A Deep Spatial and Temporal Aggregation Framework for Video-Based Facial Expression Recognition. *IEEE Access* **2019**, *7*, 48807–48815, doi:10.1109/ACCESS.2019.2907271.
151. Ghazi, M.M.; Ekenel, H.K. Automatic emotion recognition in the wild using an ensemble of static and dynamic representations. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; ACM: New York, NY, USA, 2016; pp. 514–521, doi:10.1145/2993148.2997634.
152. Almaev, T.R.; Yuce, A.; Ghitulescu, A.; Valstar, M.F. Distribution-based iterative pairwise classification of emotions in the wild using lgbp-tofp. In Proceedings of the 15th ACM on International Conference On Multimodal Interaction, Sydney, NSW, Australia, 9–13 December 2013; ACM: New York, NY, USA, 2013; pp. 535–542, doi:10.1145/2522848.2531742.
153. Gehrig, T.; Ekenel, H.K. Why is facial expression analysis in the wild challenging? In Proceedings of the 2013 on Emotion Recognition in the Wild Challenge and Workshop, Sydney, Australia, 9 December 2013; ACM: New York, NY, USA 2013; pp. 9–16, doi:10.1145/2531923.2531924.
154. Rázuri, J.G. Decision-making content of an agent affected by emotional feedback provided by capture of human's emotions through a Bimodal System. 2015. Available online: https://pdfs.semanticscholar.org/111c/55156dac0e7b31a13e80ca6a4534cd962174.pdf?_ga=2.192627626.1409604446.1572417099-1535876467.1565229560 (accessed on 1 October 2019).
155. Rashid, M.; Abu-Bakar, S.; Mokji, M. Human emotion recognition from videos using spatio-temporal and audio features. *Vis. Comput.* **2013**, *29*, 1269–1275, doi:10.1007/s00371-012-0768-y.
156. Bejani, M.; Gharavian, D.; Charkari, N.M. Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks. *Neural Comput. Appl.* **2014**, *24*, 399–412, doi:10.1007/s00521-012-1228-3.
157. Paleari, M.; Huet, B.; Chellali, R. Towards multimodal emotion recognition: A new approach. In Proceedings of the ACM International Conference on Image and Video Retrieval, Xi'an, China, 5–7 July 2010; ACM: New York, NY, USA, 2010; pp. 174–181, doi:10.1145/1816041.1816069.
158. Liu, C.; Tang, T.; Lv, K.; Wang, M. Multi-feature based emotion recognition for video clips. In Proceedings of the 2018 on International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; ACM: New York, NY, USA, 2018; pp. 630–634, doi:10.1145/3242969.3264989.
159. Mansoorizadeh, M.; Charkari, N.M. Bimodal person-dependent emotion recognition comparison of feature level and decision level information fusion. In Proceedings of the 1st International Conference on Pervasive Technologies Related to Assistive Environments, Athens, Greece, 16–18 July 2008; ACM: New York, NY, USA, 2008, p. 90, doi:10.1145/1389586.1389688.
160. Ding, W.; Xu, M.; Huang, D.; Lin, W.; Dong, M.; Yu, X.; Li, H. Audio and face video emotion recognition in the wild using deep neural networks and small datasets. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; ACM: New York, NY, USA, 2016; pp. 506–513, doi:10.1145/2993148.2997637.
161. Yao, A.; Shao, J.; Ma, N.; Chen, Y. Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; ACM: New York, NY, USA, 2015; pp. 451–458, doi:10.1145/2818346.2830585.
162. Kaya, H.; G'urpinar, F.; Afshar, S.; Salah, A.A. Contrasting and combining least squares based learners for emotion recognition in the wild. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; ACM: New York, NY, USA, 2015; pp. 459–466, doi:10.1145/2818346.2830588.
163. Ebrahimi Kahou, S.; Michalski, V.; Konda, K.; Memisevic, R.; Pal, C. Recurrent neural networks for emotion recognition in video. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; ACM: New York, NY, USA, 2015; pp. 467–474, doi:10.1145/2818346.2830596.
164. Pini, S.; Ahmed, O.B.; Cornia, M.; Baraldi, L.; Cucchiara, R.; Huet, B. Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; ACM: New York, NY, USA, 2017; pp. 536–543, doi:10.1145/3136755.3143006.

165. Gideon, J.; Zhang, B.; Aldeneh, Z.; Kim, Y.; Khorram, S.; Le, D.; Provost, E.M. Wild wild emotion: A multimodal ensemble approach. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; ACM: New York, NY, USA, 2016; pp. 501–505, doi:10.1145/2993148.2997626.
166. Chen, J.; Chen, Z.; Chi, Z.; Fu, H. Emotion recognition in the wild with feature fusion and multiple kernel learning. In Proceedings of the 16th International Conference on Multimodal Interaction, Istanbul, Turkey, 12–16 November 2014; ACM: New York, NY, USA, 2014; pp. 508–513, doi:10.1145/2663204.2666277.
167. Paleari, M.L.; Singh, V.; Huet, B.; Jain, R. Toward environment-to-environment (E2E) affective sensitive communication systems. In Proceedings of the First, ACM International Workshop on Multimedia Technologies for Distance Learning, Beijing, China, 19–24 October 2009; ACM: New York, NY, USA, 2009; pp. 19–26.
168. Sidorov, M.; Minker, W. Emotion recognition in real-world conditions with acoustic and visual features. In Proceedings of the 16th International Conference on Multimodal Interaction, Istanbul, Turkey, 12–16 November 2014; ACM: New York, NY, USA, 2014; pp. 521–524, doi:10.1145/2663204.2666279.
169. Kahou, S.E.; Pal, C.; Bouthillier, X.; Froumenty, P.; Gülçehre, Ç.; Memisevic, R.; Vincent, P.; Courville, A.; Bengio, Y.; Ferrari, R.C.; et al. Combining modality specific deep neural networks for emotion recognition in video. In Proceedings of the 15th ACM on International Conference On Multimodal Interaction, Sydney, Australia, 9–13 Decemembr 2013; ACM: New York, NY, USA, 2013; pp. 543–550, doi:10.1145/2522848.2531745.
170. Krishna, T.; Rai, A.; Bansal, S.; Khandelwal, S.; Gupta, S.; Goyal, D. Emotion recognition using facial and audio features. In Proceedings of the 15th ACM on International Conference On Multimodal Interaction, Sydney, Australia, 9–13 December 2010; ACM: New York, NY, USA, 2013; pp. 557–564, doi:10.1145/2522848.2531746.
171. Wang, H.; Huang, H.; Hu, Y.; Anderson, M.; Rollins, P.; Makedon, F. Emotion detection via discriminative kernel method. In Proceedings of the 3rd International Conference on Pervasive Technologies Related to Assistive Environments, Corfu, Greece, 26–29 June 2018; ACM: New York, NY, USA, 2010, p. 7, doi:10.1145/1839294.1839303.
172. Nicolaou, M.A.; Gunes, H.; Pantic, M. A multi-layer hybrid framework for dimensional emotion classification. In Proceedings of the 19th ACM International Conference on Multimedia, Scottsdale, AZ, USA, November 28–December 01 2011; ACM: New York, NY, USA, 2011; pp. 933–936, doi:10.1145/2072298.2071906.
173. Chao, L.; Tao, J.; Yang, M.; Li, Y.; Wen, Z. Multi-scale temporal modeling for dimensional emotion recognition in video. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Orlando, FL, USA, 7 November 2014; ACM: New York, NY, USA, 2014; pp. 11–18, doi:10.1145/2661806.2661811.
174. Meng, H.; Bianchi-Berthouze, N. Affective state level recognition in naturalistic facial and vocal expressions. *IEEE Trans. Cybern.* **2013**, *44*, 315–328, doi:10.1109/TCYB.2013.2253768.
175. Song, Y.; Morency, L.P.; Davis, R. Learning a sparse codebook of facial and body microexpressions for emotion recognition. In Proceedings of the 15th ACM on International Conference On Multimodal Interaction, Sydney, Australia, 9–13 December 2013; ACM: New York, NY, USA, 2013; pp. 237–244, doi:10.1145/2522848.2522851.
176. Meng, H.; Bianchi-Berthouze, N.; Deng, Y.; Cheng, J.; Cosmas, J.P. Time-delay neural network for continuous emotional dimension prediction from facial expression sequences. *IEEE Trans. Cybern.* **2015**, *46*, 916–929, doi:10.1109/TCYB.2015.2418092.
177. Liu, M.; Wang, R.; Li, S.; Shan, S.; Huang, Z.; Chen, X. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In Proceedings of the 16th International Conference on Multimodal Interaction, Istanbul, Turkey, 12–16 November 2014; ACM: New York, NY, USA, 2014; pp. 494–501, doi:10.1145/2663204.2666274.
178. Fan, Y.; Lu, X.; Li, D.; Liu, Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; ACM: New York, NY, USA, 2016; pp. 445–450, doi:10.1145/2993148.2997632.
179. 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; ACM: New York, NY, USA, 2017; pp. 553–560, doi:10.1145/3136755.3143009.

- 180. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep face recognition. *BMVC*, **2015**, *1*, 6.
- 181. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).