Facial Expression Recognition

Thi Huyen Cao

University of Applied Science Hamburg
Berliner Tor 5, 20099 Hamburg, Germany

Abstract. Facial expression is a very important source of information during communicating. It helps communicators identify the current emotional state and react to it in an appropriated manner. For decades there are many works that have been trying to encode this information automatically. Facial Expression Recognition (FER) refers to the task of detecting human emotion through the facial expression. It is one of the most interesting and challenging researches in Computer Vision field. The aim of this paper is to give a comprehensive overview about FER, including different approaches to emotion, a summary of current available datasets, a brief introduction to common techniques as well as challenges and opportunities of FER.

Keywords: Facial Expression Recognition (FER) · Machine Learning · Deep Learning

1 Introduction

Emotion is one of the most important channels in human communication system. Humans express their emotion by facial expression, speech or body language naturally despite their age, culture, personal background or gender. Facial expression is considered as the source of most information when it comes to Emotion Recognition. Therefore, the 2 terms Facial Expression Recognition and Emotion Recognition are often mentioned together or sometimes implied as one. While it seems very simple for human to express, interpret and interact with facial expression, it is not for machine. Under machine's perspective, an image is nothing more than a combination of pixels in form of number. Finding patterns and features from those numerical values to define an emotional state is hence very difficult. In recent years, there is considerable growing interest in Facial Expression Recognition (FER) due to its wide variety of application fields including medicine, data analysis, human-computer interaction, robotics, education as well as car and film industry. While most works in the early stages only focused on static images in laboratory environment, more and more studies on videos and real world application scenes are published lately. Some basic major issues of unconstrained real world environment are face occlusion and pose variation. Based on the assumption of some basic universal emotions [1], the most popular task of FER is classification, which means assigning each input to a class of an emotion among n classes. While this approach is simple and interesting, it fails to take into account the complexity of human emotion system. Until

now, there are continuous discussions about different approaches to emotion. As Machine Learning matures, it becomes the state of the art of a lot of challenging tasks. FER is not an exception. [2] addressed different neural networks and training strategies, which are designed for FER.

In this paper, the author introduces a comprehensive overview of FER. The rest of the paper is organized as follows. Some main approaches to FER are presented in section 2. Section 3 describes frequently used datasets. Standard pipeline of FER is provided in section 4. Section 5 investigates the challenges of FER including face pose variation, collecting data, ground truth, context, along with others. Many application scenes are also discussed in this section. Some conclusions are drawn in the last section.

2 Approaches

Again, among different factors, facial expression is the most important one to determine the state of emotion. [3] clears the distinction by referring facial expression as the "signal" and emotion as the "message". Different messages can be interpreted from the same signal in different contexts or by different people. While a smile is associated directly to "happiness" in most cases, it is expressed very common as "anxiety" of people in uncomfortable situation. In 1978, Ekman proposed a system for encoding the "signal" without involving the "message" called Facial Action Coding System (FACS). Each movement of specific regions such as eyes, mouth, cheeks and so on is called Action Unit (AU). An example is showed in Figure 1.

Fig. 1. Two expressive images and the list of active AU (together with their physical meaning) [5]

These AUs are then used for recognizing the emotion. The occurrence of a combination of specified AUs can be interpreted as a specific emotion. [6] demonstrates some examples in table 1.

The system of 6 emotions mentioned in table 1 is also one of the most popular representative of **categorical approach**, also known as discrete emotion approach. This approach refers that there are a number of distinguished emotions.

Table 1. Emotional FACS for 6 basic emotions [6]

Emotion	Action Units
Happiness	6,12
Sadness	1,4,15
Surprise	1,2,5B
Fear	1, 2, 4, 5, 7, 20, 26
Anger	4, 5, 7, 23
Disgust	9, 15, 16

Darwin, father of the theory of evolution, is often the first one to be mentioned when it comes to the hypothesis of universal facial behaviour [7] although there are more works from other researchers before and after him. Ekman and Friesen [1] supported this theory by designing, executing and evaluating an experiment, in which they told different groups of people stories and let them pick one out of 3 faces, which is mostly appropriate to the story. The results strengthened the hypothesis that particular emotions are indeed universal. They came up with 6 basic universal emotions which are {anger, disgust, fear, happiness, sadness, surprise}. Following the same concept of discrete emotion, system of 7, 8 basic emotions, the wheel of emotion etc. are addressed in [8,9]. In contrast to discrete approach, **dimensional approach**, also known as continuous emotion approach, do not distinguish among different emotions but consider them as points in a continuous space. This approach emphasizes that there are factors, which exist in every single emotion and therefore each emotion can be presented by a point in a n-factors coordinate system. E.g. 2 dimensions system of arousal and valence, 3 dimensions of pleasantness-unpleasantness, attention-rejection and level of activation. It also illustrates the complexity of human emotion better than categorical approach. There are continuous discussions around these 2 approaches. Nevertheless, categorical approach is often chosen over dimensional approach for the sake of simplicity, application context and goals.

3 Datasets

Data with good quality is one of the most essential factors to train a robust automatic FER system, especially for approaches using supervised learning algorithm. There are a number of publicly available datasets, which are frequently used for evaluating FER systems. These datasets differ from the following points:

- Environment: laboratory-controlled, in the wild (web, movies and real world application)
- Approach: basic emotions, value in arousal-valence scale, Action Units (AUs)
- Size: total size, number of objects, number of emotions, resolution, frame rate
- Type: static images, frame sequences (video)
- Color: gray, RGB
- Dimension: 2D, 3D

- Capture: posed, spontaneous
- Bias: object gender, ethics, origin

Most of the datasets are created in the laboratory with good set up of light condition, contrast and position, contains 2D RGB images with extreme emotions, uses the categorical approach to annotate these images with a number of basic emotions. Each category of dataset has advantages and disadvantages. E.g. posed datasets are easy to capture high intensity of emotion but fails in describing natural human behavior (expression of emotion in a range of intensity). Depending on the specific tasks, a corresponding dataset or a combination of datasets should be used for training. Below is the brief introduction of some most commonly used datasets from the combination of some categories aboved. A further read can be found in [2].

In laboratory

1. CK+ [10]: CK is one of the very first comprehensive databases for FER. It was publicly released in 2000. An extension followed in 2010 with improvement in labeling, increase in size etc. CK+ is among the most used laboratory-controlled FER datasets. It contains 593 video sequences from 10 to 60 frames with resolution of 640x490 from 123 objects. It provides FACS coding as well as validated emotion label. Among all, 327 sequences from 118 objects are annotated with an emotion in 7 basic expression labels of {sadness, surprise, happiness, fear, anger, contempt, disgust}. CK+ has posed as well as spontaneous facial expressions.
2. MMI [11]: MMI dataset is created in laboratory and named after their creators **M**aja Pantic, **M**ichel Valstar and **I**oannis Patras. It includes more than 1500 samples of both static and image sequences in frontal view with resolution of 720x576. Similar to CK+, it provides posed and spontaneous expressions with AU and emotion label. The labels are based on 6 basic emotions like CK+ except contempt. MMI exposes objects of different gender with varying ethnic background, different challenging conditions such as wearing accessories, mustache etc. Nonetheless, like CK+ and many other laboratory-controlled datasets, it does not investigate occluded faces (by hand or a long beard). Most importantly, MMI is designed and delivered as a web-based application and therefore is considered as one of the most easily accessible and searchable resources for FER.

In the wild

1. EmotioNet [12]: EmotioNet database comes originally from EmotioNet Challenge, which has been occurred annually since 2016. It provides a large number of images collection ($\sim$ 1 million) from the internet for evaluating the computer vision algorithms in automatic FER. The term of the challenge changes year by year. At the beginning, it contains 2 tasks: automatic AU detection and automatic emotion recognition. The database is only available for research purposes and can be downloaded from [13].

2. AFEW [14]: The Acted Facial Expressions in the Wild (AFEW) serves as the database for the Emotion Recognition In The Wild Challenge (EmotiW) since 2013. It contains video clips from movies, in which it exposes spontaneous expressions, various challenging conditions including head position, occlusion and illustration. Samples are labeled with 6 basic emotions plus neutral. The database is continuously updated and reality TV shows data has been added lately.

In many datasets there are no explicit split among training set, test set and development set. Hence it is difficult to evaluate the performance of different algorithms on the same dataset. In addition, it's very important to split the data carefully so that the model does not learn by heart but learn to generalize.

4 Pipeline

The standard algorithmic pipeline for FER consists of 3 major steps, as showed in Figure 2 whereas Deep Learning skips the second step since neural networks are capable of learning features directly from raw data.

Fig. 2. Standard pipeline of FER

Each component will be investigated concretely in the following in term of its aims, challenges, state of the art and best practice.

4.1 Preprocessing

This step aims to align and normalize the visual semantic information from faces as there are many possible noises and variation from the original images such as different backgrounds, illuminations, occlusion and head poses. Some common preprocessing steps are face detection, face alignment, illumination & pose normalization and data augmentation.

– **Face detection**: Every face analysis begins with face detection. According to [15], there are different approaches to look for faces such as using typical skin color to find face segments, finding faces by motion, using Edge-Orientation Matching, Hausdorff Distance, using Cascade of classifier, using Histogram of Oriented Gradients (HOG) or Deep Learning. A classical face detector is Viola & Jones Haar Cascade (V & J), which was first proposed in 2001 [16]. V & J face detector is robust, computationally simple and works best with frontal faces. It can be used easily with some lines of

code from OpenCV [1]. In addition, many Deep Learning face detectors have been exploited recently and has gained remarkable results such as Multi-task Convolutional Neural Network (MTCNN), Tasks-Constrained Deep Convolutional Network (TCDCN), RetinaFace, along with others. Deep Learning has become the state of the art of face detection in term of high speed and accuracy. Best practice in challenging environments is to use a combination of different detectors. For detecting faces in video, tracker is commonly used to assist the detector thank to its character of maintaining temporal changes. In general, tracking is more efficient than frame-to-frame detecting.

- **Face alignment or facial landmarking**: [5] pointed out that localization of fine-grained inner facial structures such as lip, eyes, nose etc. is a great help at direct extraction of geometric features as well as more powerful local features. Therefore, this is also one of the very popular and advisable preprocessing steps. [2] investigates many facial landmark detection algorithms in term of efficiency and accuracy. Its result shows that supervised descent method (SDM), a discriminative model which is implemented as a part of the public-available software package IntraFace (IF) achieves the state of the art due to its simple implementation, low computation cost and robustness to images and videos in the wild. Deep neural networks such as MTCNN and TCDCN can be counted as good alternatives. Like face detection, facial landmark tracker is widely-used in videos to achieve better performance and deal with occlusions better.
- **Nuisance factor reduction**: In reality scenarios, faces are recorded under unconstrained environments, and therefore contains many noisy factors such as too much light, too low contrast, occlusion, unreasonable position along with others. 2 common techniques to reduce nuisance factors are illumination & pose normalization. Illumination normalization refers to the task of dealing with illumination variation. Isotropic diffusion (IS)-based normalization, discrete cosine transform (DCT)-based normalization and difference of Gaussian (DoG) are some methods to gain illumination invariant. [2] suggests that a combination of illumination normalization and histogram equalization results better FER than illumination normalization alone. Pose normalization is about yielding the frontal view for detected faces. [2] summarizes 2 approaches to this task, which are using (1) syntheses to recreate/convert the face into frontal face based on found facial landmarks and (2) applying generative adversarial network (GAN).
- **Data augmentation**: Supervised Machine Learning (ML) algorithms require especially a lot of data to be able to generalize. In addition, the data needs to have good quality and correct label. Collecting enough data for training such algorithm is becoming less difficult thank to the rise of the internet. Still it's a resource-consuming process in term of money, time and human effort. Data augmentation is the best help while dealing with data problem. The art of data augmentation is to "invent" more data from the original data by a variety of transformation methods such as shifting, ro-

[1] A popular open source library for Computer Vision and Machine Learning

tating, adding noises, equalizing histogram among others. In addition to increasing the size of data, it also increases the diversity of data by generating different contexts, backgrounds or positions. There are 2 types of data augmentation: online and offline. While offline means augmenting the data completely before training, it is used to leading to exploding size of the database at big original dataset. In FER, online data augmentation is commonly used. It generates new data during the training process. Many libraries like Keras offer methods to do online data augmentation in some lines of code. In the last few years, there is considerable growing interest in using generative adversarial network (GAN), a Deep Learning based technology to generate more facial expressions as well as various face poses.

4.2 Feature extraction

There are 2 types of feature: handcrafted and learned feature. With Deep Learning, the neural networks are capable of learning the necessary pattern (learned feature) from raw data, no feature extraction task is needed. Other Machine Learning algorithms might require prepared features for training. [5] mentions 3 categories of feature which can be extracted in advance: appearance, geometric and motion feature. As the names indicate, the first one encodes pixel intensity information, the second one is based on the facial landmark location and the last one captures the changes among frames. Some examples of each category are showed in the following table.

Table 2. Examples of handcrafted feature

feature category	examples
appearance features	Local Binary Pattern (LBP), Gabor, Histogram of Oriented Gradients (HOG), Three Orthogonal Planes (TOP), Gaussian Laguerre (GL), Weber Local Descriptor (WLD), Discrete Contourlet Transform (DCT) [17]
geometric features	distance among landmarks, angle between neutral face and input face, angle formed by the segments joining three landmarks [5], Local Curvelet Transform (LCT) [17]
motion features	Motion History Images [18]

Depending on the input images as well as the defined goals, corresponding features should be extracted. In many cases, hybrid feature extraction (a combination of features from different categories) helps get better performance at training.

4.3 Feature learning

Feature learning is the final stage of an automatic FER system. The learned features are mostly used for classification of n discrete emotions. Nevertheless,

there are also regression tasks like pain estimation and so on. Machine Learning algorithms are generally used in this final step including K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Deep Learning (DL), Hidden Markov Model (HMM), Decision Tree (DT) etc. In the following there is a brief introduction of some most efficient algorithms for FER whereas Deep Learning has achieved the state of the art. A further read about each algorithm mentioned above as well as some comparisons among them can be found in [2,17].

- **Support Vector Machine (SVM)**: SVM is originally designed for binary classification. The goal of this algorithm is to find a maximum margin hyperplane between 2 classes. While in linear separated datasets, a line will handle the split, kernel trick is used in nonlinear separable datasets by creating a new dimension. Multi-class SVM is possible and is simply a combination of many binary SVM classifiers.
  - one vs the rest: train k binary classifiers for k classes. Each classifier determines between its own class and the rest
  - one vs one: train a binary classifier for each pair of classes

  Many works has exploited the performance of SVM classifier for FER combined with Principal Component Analysis (PCA) algorithms, Gabor and Local Binary Pattern (LBP) features [19,20]. The results were absolutely better than the random baseline, achieved in some datasets an accuracy of over 70% .
- **Deep Learning (DL)**: DL has recently achieved the state of the art performance of many applications including FER. The art of DL is using deep neural networks, which are inspired from the human brain to learn the features, patterns and rules from raw data and therefore offers an end-to-end training process. [2] did a comprehensive survey about deep FER systems with static images and dynamic image sequences. They examined different neural network architectures and corresponding training strategies. The results showed that transfer learning is the mainstream in deep FER to solve problems with insufficient training data and overfitting. Pre-trained neural networks for object detection have learned many basic features of identifying edges, curves etc. and can be fine-tuned for FER. [2] emphasized the efficiency of fine-tuning pre-trained networks in multiple stages. It also pointed out an alternative to train the model from scratch by using well-designed auxiliaries and blocks to enhance the learning capacity. These designs force the model to focus more on the relevant information. Another suggested way to train a good FER system is to ensemble multiple networks. [2] mentioned 2 important keys to be considered when implementing such a network ensemble: (1) sufficient diversity of the networks to ensure complementarity (2) an appropriate ensemble method. In reality, humans express their emotions in a dynamic process. Building such a FER system on image sequences is therefore necessary for many applications. Frame aggregation is a simple technique to combine the learned feature (feature-level aggregation) or prediction probability (decision-level aggregation) of each frame from the whole sequence. Unfortunately, this technique does not take the intensity

of expression as well as the temporal factor into consideration, which are 2 natural characteristics of emotion expression. To have a better performance of frame aggregation, the intensity of each frame might need to be provided in advance and weighted during aggregating. Deep spatio-temporal neural networks handle this problem in a more elegant way. These networks are cascaded from different networks, each with its own advantage. For example a cascaded network of CNN as frontend layers and LSTM as backend layers is indeed able to learn both space features (how the facial expression looks like) and time features (how the facial expression is changing). According to [2], many cascaded networks have been proposed and proved their performance for FER. There are many other neural networks as well as training techniques which can not be discussed all here. A detailed description of them can be found in [2].

5 Challenges and Opportunities

A robust FER system will open opportunities for applications in a variety of fields. Researches of FER in different domains are published day by day. Below are just some representatives among many applied fields.

- Medicine: Among other tools of monitoring and analyzing human behaviors, FER will help detect depression and extreme anxiety in the early stage so that appropriated treatment can be applied in time. In addition, it is also potential to apply FER in pain intensity estimation.
- Education and entertainment: Online learning has become more and more popular nowadays because of its advantages in location and presentation. However, one of its huge disadvantages is the impersonal interaction between the platform and user. Individual-based online learning platform as well as games will be possible with FER. Depending on the current emotional state, a learning management system might offer the learner a new exercise, a difficult level or a break. Emotional trigger can be implemented in game with the same principle to have a better interaction with player.
- Data analytics: One of the most important representatives from the economy side is marketing. FER can help evaluate the effectiveness of marketing campaigns by analyzing audience interaction at a screening session or in front of camera while watching commercial for instance.
- Robotics: Social robots are required to understand and interact in an elegant level with humans. As expressing emotions is a natural characteristic of human, it needs to be minored in such robots as well. FER will be great help in implementing such features. Furthermore, FER also plays an important role in the field of human-computer interaction. With a better understanding of human facial expression, it will be able to build a more precise human-computer interface in many different applications.
- Car industry: Drowsiness detection will be able with FER. Giving the driver a warning in time if she/he is tired will help significant at reducing the number of accidents.

– Social networks: Tagging is the process of giving keywords to the content of media such as name in photos in Facebook, hashtag # for photos in Instagram or videos in Youtube. FER will help bring this process to the next level, which is providing keywords about the content as well as the emotion detected.

Nonetheless, there are many challenges for FER, especially in real-world scenarios and with limitation of computation resource (mobile applications). The following summarizes some of them and possible solutions involved.

– Data and ground truth: Training a robust FER system requires good data with good label. Collecting facial expression data is never an easy task. Images of happiness with a big smile are for example easy to find in the internet while it's very difficult to find pictures of anger, which leads to imbalanced distribution in many available datasets. Handling this imbalance is a very important point while training a FER model. Inconsistent annotation is another problem with the current datasets, where the same or identical images are labeled differently. This makes it very difficult to evaluate the performance of algorithm on different datasets. Creating ground truth is hence challenging since same facial expression can be thus interpreted differently by different annotators. It is a very time-consuming process since for the objectivity of the label, in most cases annotations from a number of people are required. One of the efficient methods to annotate facial expression in video is to use joy stick to define the emotional state. Many real-world applications require FER system to be capable of handling a level of occlusion as well as different face poses. GAN is one of the techniques which is recently widely-used to frontalize face, reconstruct occluded part, generate a plenty number of poses, facial expressions to help train such a FER system for application scenarios in the wild.
– Context and bias: Even there are after some theories a number of universal emotions, each individual expresses his/her feeling in a certain way, which is effected by many internal and external factors (movies, other people in the environment, genetics, experiences, culture etc.). To fully understanding the facial expression, context is needed in many cases.
– Multimodal affect recognition: Last but not least, understanding human behavior or even only human emotion requires encoding different channels such as electroencephalography (EEG), audio, mimics, context where facial expression is only one of them. [2] pointed out the promising result from only facial expression and emphasized the effectiveness of a multimodal system by incorporating information from other models in a high-level framework to complement information and enhance the robustness.

6 Conclusion

There are many different approaches to emotion which are discussed for years. 2 popular ones are discrete and continuous emotion. Facial expression is one out of

many other channels which contribute to the work of defining an emotion state. A lot of public available datasets of images and videos collected from laboratory and in the wild can be used for training an automatic FER system. The standard pipeline for training a FER model consists of 3 stages: preprocessing, feature extraction and feature training. Deep Learning, with its continuous development, has achieved the state of the art of both static and dynamic FER. Many works have been proposed to solve challenges of FER as well as sub tasks like face detection, face alignment etc. There is continuously growing interest in this field because of its huge potential of applications. Although there are many not-yet-solved challenges around FER, the author deeply believes that FER has reached a maturity to be implemented in real world applications.

References

1. Authors, P. Ekman and W. V. Friesen: Article title. Constants across cultures in the face and emotion, Journal of personality and social psychology , vol. 17, no. 2, pp. 124129 (1971)
2. Authors, Shan Li and Weihong Deng : Article title. Deep Facial Expression Recognition: A Survey (2018) http://arxiv.org/abs/1804.08348
3. Authors, J. F. Cohn and P. Ekman: Measuring facial actions In: Book title. The New Handbook of Methods in Nonverbal Behavior Research, Harrigan, J.A., Rosenthal, R. & Scherer, K., Eds., pp. 964. Oxford University Press (2005)
4. Authors, P. Ekman and W. V. Friesen: Article title. Facial Action Coding System: A technique for the measurement of facial movement In: Consulting Psychologists Press (1978)
5. Authors, Brais Martinez and Michel F. Valstar. Advances: Chapter title. Challenges and Opportunities in Automatic Facial Expression Recognition In: Book title. Advances in Face Detection and Facial Image Analysis (2020)
6. TU Chemitz, https://www.tu-chemnitz.de/informatik/KI/projects/facedetection/index.php. Last accessed 13 Feb 2020
7. Author, C. Darwin: Book title. The Expression of the Emotions in Man and Animals, Publisher John Murray (1872)
8. Author, D. Matsumoto. Article title. More evidence for the universality of a contempt expression In: Book title. Motivation and Emotion (1992)
9. PositivePsychology, https://positivepsychology.com/emotion-wheel/. Last accessed 14 Feb 2020
10. Authors, P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews: Article title. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in Computer Vision and Pattern Recognition Workshops (CVPRW) In: IEEE Com- puter Society Conference pp. 94–101 (2010)
11. Authors, M. Pantic, M. Valstar, R. Rademaker, and L. Maat: Article title. Web-based database for facial expression analysis In: IEEE International Conference (2005)
12. Authors, C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, Qianli Feng, Yan Wang, Aleix M. Martinez: Article title. EmotioNet Challenge: Recognition of facial expressions of emotion in the wild (2017) https://arxiv.org/pdf/1703.01210.pdf

13. Computational Biology and Cognitive Science Lab of The Ohio State University, http://cbcsl.ece.ohio-state.edu/dbform_emotionet.html. Last accessed 16 Feb 2020
14. Authors, A. Dhall, R. Goecke, S. Lucey, and T. Gedeon: Article title. Collect-ing large, richly annotated facial-expression databases from movies (2012)
15. Face Detection and Recognition Homepage by Dr. Robert Frischholz, https://facedetection.com/algorithms/. Last accessed 14 Feb 2020
16. Authors, Paul Viola and Michael Jones: Article title. Rapid Object Detection using a Boosted Cascade of Simple Features (2001)
17. Authors, I.Michael Revina and W.R. Sam Emmanuel: Article title. A Survey on Human Face Expression Recognition Techniques In: IEEE Journal (2018)
18. Authors, Sander Koelstra, Maja Pantic, Ioannis (Yannis) Patras: Article title. A dynamic texture based approach to recognition of facial actions and their temporal models In: IEEE Journal (2010)
19. Authors, Muzammil Abdulrahman and Alaa Eleyan: Contribution title. Facial expression recognition using Support Vector Machines In: 23nd Signal Processing and Communications Applications Conference (SIU), pp. 276–279 (2015)
20. Authors, Vasanth P.C.Nataraj K.R. : Article title. Facial Expression Recognition Using SVM Classifier In: Indonesian Journal of Electrical Engineering and Informatics (IJEEI) (2015)