**INTRODUCTION**

F ACIAL expression is one of the most powerful, natural and universal signals for human beings to convey their emotional states and intentions [1], [2]. Numerous studies have been conducted on automatic facial expression analysis because of its practical importance in sociable robotics, medical treatment, driver fatigue surveillance, and many other human-computer interaction systems. In the field of computer vision and machine learning, various facial expression recognition (FER) systems have been explored to encode expression information from facial representations. As early as the twentieth century, Ekman and Friesen [3] defined six basic emotions based on cross-culture study [4], which indicated that humans perceive certain basic emotions in the same way regardless of culture. These prototypical facial expressions are anger, disgust, fear, happiness, sadness, and surprise. Contempt was subsequently added as one of the basic emotions [5]. Recently, advanced research on neuroscience and psychology argued that the model of six basic emotions are culture-specific and not universal [6].

Although the affect model based on basic emotions is limited in the ability to represent the complexity and subtlety of our daily affective displays [7], [8], [9], and other emotion description models, such as the Facial Action Coding System (FACS) [10] and the continuous model using affect dimensions [11], are considered to represent a wider range of emotions, the categorical model that describes emotions in terms of discrete basic emotions is still the most popular perspective for FER, due to its pioneering investigations along with the direct and intuitive definition of facial expressions. And in this survey, we will limit our discussion on FER based on the categorical model.

FER systems can be divided into two main categories according to the feature representations: static image FER and dynamic sequence FER. In static-based methods [12], [13], [14], the feature representation is encoded with only spatial information from the current single image, whereas dynamic-based methods [15], [16], [17] consider the temporal relation among contiguous frames in the input facial expression sequence. Based on these two visionbased methods, other modalities, such as audio and physiological channels, have also been used in multimodal systems [18] to assist the recognition of expression.

The majority of the traditional methods have used handcrafted features or shallow learning (e.g., local binary patterns (LBP) [12], LBP on three orthogonal planes (LBP-TOP) [15], non-negative matrix factorization (NMF) [19] and sparse learning [20]) for FER. However, since 2013, emotion recognition competitions such as FER2013 [21] and Emotion Recognition in the Wild (EmotiW) [22], [23], [24] have collected relatively sufficient training data from challenging real-world scenarios, which implicitly promote the transition of FER from lab-controlled to in-the-wild settings. In the meanwhile, due to the dramatically increased chip processing abilities (e.g., GPU units) and well-designed network architecture, studies in various fields have begun to transfer to deep learning methods, which have achieved the state-of-the-art recognition accuracy and exceeded previous results by a large margin (e.g., [25], [26], [27], [28]). Likewise, given with more effective training data of facial expression, deep learning techniques have increasingly been implemented to handle the challenging factors for emotion recognition in the wild. Figure 1 illustrates this evolution on FER in the aspect of algorithms and datasets.

Exhaustive surveys on automatic expression analysis have been published in recent years [7], [8], [29], [30]. These surveys have established a set of standard algorithmic pipelines for FER. However, they focus on traditional methods, and deep learning has rarely been reviewed. Very recently, FER based on deep learning has been surveyed in [31], which is a brief review without introductions on FER datasets and technical details on deep FER. Therefore, in this paper, we make a systematic research on deep learning for FER tasks based on both static images and videos (image sequences). We aim to give a newcomer to this filed an overview of the systematic framework and prime skills for deep

Despite the powerful feature learning ability of deep learning, problems remain when applied to FER. First, deep neural networks require a large amount of training data to avoid overfitting. However, the existing facial expression databases are not sufficient to train the well-known neural network with deep architecture that achieved the most promising results in object recognition tasks. Additionally, high inter-subject variations exist due to different personal attributes, such as age, gender, ethnic backgrounds and level of expressiveness [32]. In addition to subject identity bias, variations in pose, illumination and occlusions are common in unconstrained facial expression scenarios. These factors are nonlinearly coupled with facial expressions and therefore strengthen the requirement of deep networks to address the large intra-class variability and to learn effective expression-specific representations.