Deep Learning for Human Affect Recognition: Insights and New Developments

Philipp V. Rouast, *Student Member, IEEE,* Marc T. P. Adam, and Raymond Chiong, *Senior Member, IEEE*

Abstract—Automatic human affect recognition is a key step towards more natural human-computer interaction. Recent trends include recognition in the wild using a fusion of audiovisual and physiological sensors, a challenging setting for conventional machine learning algorithms. Since 2010, novel deep learning algorithms have been applied increasingly in this field. In this paper, we review the literature on human affect recognition between 2010 and 2017, with a special focus on approaches using deep neural networks. By classifying a total of 950 studies according to their usage of shallow or deep architectures, we are able to show a trend towards deep learning. Reviewing a subset of 233 studies that employ deep neural networks, we comprehensively quantify their applications in this field. We find that deep learning is used for learning of (i) spatial feature representations, (ii) temporal feature representations, and (iii) joint feature representations for multimodal sensor data. Exemplary state-of-the-art architectures illustrate the progress. Our findings show the role deep architectures will play in human affect recognition, and can serve as a reference point for researchers working on related applications.

Index Terms—Affect recognition, Deep learning, Emotion recognition, Human-computer interaction.

◆

1 INTRODUCTION

HUMAN-COMPUTER INTERACTION (HCI) is seeing a gradual shift from the computer-centered approach of the past to a more user-centered approach. Commercial success has made user-friendly input methods, portable devices and multi-sensor availability a new standard in personal computing. Despite the progress, it has been argued that HCI is still lacking a central element of human-human interaction: The communication of information through affective display [1]. The term *affective computing* was coined by Rosalind Picard in 1995 [2], inspired by findings from neuroscience, psychology, and cognitive science highlighting the important role affect plays in intelligent behavior. It encompasses efforts to (i) automatically recognize human affect, and (ii) generate corresponding responses by the computer, providing a richer context for HCI outcomes. Applications range from education [3] and health care [4] to entertainment [5] and embodied agents [6].

In human interactions, a significant amount of information is not communicated explicitly, but through the way we speak, our facial expressions, gestures, and other means. Initial research on affect recognition[1] focused mainly on unimodal approaches, with facial expression recognition (FER) and speech emotion recognition (SER) gaining most attention and highest accuracies [7]. Psychophysiological measures were also shown to contain information about affective states [8]. Public database availability has improved since the 2000s [9] and multimodal sensor combination was found to improve recognition accuracy and robustness [10]. As good performance was achieved on posed databases,

the focus started to shift towards more realistic, spontaneous displays of affective behavior [11]. These trends are exemplified in the annual competitions *Emotion Recognition in the Wild* (EmotiW) [12] and *Audio Video Emotion Challenge* (AVEC) [13]. Since 2010, *deep learning* methods have been applied to affect recognition problems across multiple modalities and led to improvements in accuracy, including winning performances at EmotiW [14], [15], [16] and AVEC [17], [18], [19].

In this paper, we review the state of research on affect recognition using deep learning. We start by giving an introduction to deep learning, the reasoning behind its application in artificial intelligence (AI), and to the most relevant architectures. We then break down the challenges faced in affect recognition research, and outline why models incorporating deep learning are useful in meeting these. Finally, we go on to discuss the directions research has taken since 2010 and in what way deep learning methods are utilized. Our contributions to the field are:

- By conducting a comprehensive literature search, and classification of a total of 950 studies, we are able to identify and measure a trend towards use of deep neural networks for affect recognition;
- reviewing 233 studies that employ deep learning, we identify the main application areas as being the learning of (i) *spatial feature representations*, (ii) *temporal feature representations*, and (iii) *joint feature representations* for multimodal sensor data;
- we discuss and give exemplary illustrations of how deep neural networks are applied on visual, auditory, and physiological sensor data;
- we provide an overview of the most relevant databases by quantifying their use across 233 studies, including large databases established since 2016; and
- we discuss open issues and research directions.

- *P. V. Rouast, M. T. P. Adam and R. Chiong are with the School of Electrical Engineering and Computing, University of Newcastle, Callaghan, NSW 2308, Australia.*
  E-mail: philipp.rouast@uon.edu.au

1. Unless stated otherwise, in this article we use the term affect recognition to refer to human affect recognition.

2 DEEP LEARNING FOR AFFECT RECOGNITION

Automatic affect recognition relies on abilities commonly expected of AI, such as visual and auditory perception [20]. A class of methods known as *machine learning* has turned out to be effective in handling the desired perception tasks. By enabling computers to learn directly from examples, these algorithms overcome the need to provide an explicit model.

Although perception tasks such as visual and auditory affect recognition seem intuitive to humans, some characteristics rooted in their real-world origin make them hard problems to solve: They require understanding of problems characterized by highly varying functions in terms of the input; another common characteristic is the high dimensionality of examples in the form of images and audio files.

In this context, a common challenge for traditional machine learning approaches is the *curse of dimensionality*, a phenomenon where the higher the number of dimensions used to represent the data, the less effective conventional computational and statistical methods become [21]. With a very high number of dimensions, it becomes increasingly difficult to comprehensively sample all possible combinations, resulting in vast unexplored regions in the feature space. To circumvent this problem, a straightforward and widely used solution is to project the high-dimensional data into a lower-dimensional space through approaches such as feature selection. Machine learning algorithms with so-called *shallow* architectures, such as kernel methods and single-layer neural networks, can then be efficiently applied for modeling purposes. However, when considering computational and statistical efficiency as well as human involvement, it has been suggested that shallow architectures may not be the most efficient way to approach challenging learning problems such as affect recognition [20]. Hence, in 2010 researchers have started to explore the application of *deep* architectures for affect recognition.

2.1 Deep learning

To distinguish between *shallow* and *deep* machine learning models, one can think of their architectures as subsequent layers of hierarchical computation. We will use the notion of architecture *depth* to denote the number of computational layers in an architecture[2]. Traditional learning algorithms can generally be represented as two layers of computation, where the first layer consists of template matchers or simple trainable basis functions, and the second layer is a weighted sum [20]. This is why we talk of them as *shallow* architectures, when comparing them with *deep* neural networks, which consist of three or more layers. Although there is no universally agreed upon rule to determine depth or distinguish between shallow and deep architectures [22, ch.1], a cutoff of three or more layers is commonly used [23], [24], [25].

2.1.1 Deep neural networks

The design of deep neural networks (DNNs) is loosely inspired by biological neural networks. A typical example is the deep feedforward network (or multi-layer perceptron). It consists of multiple layers of processing units ("neurons"):

An input layer, multiple hidden layers, and an output layer. Units in adjacent layers can have weighted connections. We speak of *fully-connected* DNNs if there are connections between all pairs of units in adjacent layers. Information in the network flows forward through these connections, each unit computing its *activation* as a function of its inputs. Units in hidden layers introduce a nonlinearity in the process. By adjusting the connection weights, a DNN can effectively learn a *feature representation* of its input data in each layer's unit activations. A well-trained DNN learns a deep hierarchy of *distributed representations*[3]. This enables the network to learn very expressive representations capturing a large number of possible input configurations [26].

Some of the key advantages of deep architectures are derived from their depth. Increasing depth promotes re-use of learned features [27]. On a related note, the deep hierarchy of feature representations allows learning at different levels of abstraction building on top of each other. Here, higher levels of abstraction are generally associated with invariance to local changes of the input [27].

While the theoretical advantages of such deep architectures were known for some time, progress was held back by the difficulty of training them [25]. An initial breakthrough in 2006 [23] showed that the so-called *vanishing gradient problem* in training DNNs can be overcome by unsupervised pre-training. Multiple strategies of alleviating the problem are known today, including (i) architectures unaffected by it [28], [29], (ii) improved optimizers [30], (iii) certain training and design choices [31], [32], [33], and (iv) use of powerful computing systems, especially GPU-based. Three interrelated factors drive the continued success of DNNs:

- *Increased learning capacity*[4]. A central theme in the evolution of deep models has been a link between better generalization ability and increased number of parameters, especially when growing models *deeper* rather than *wider* [22, ch.6.4]. This applies as long as training is feasible and the dataset is large enough to take advantage of the architecture [29].
- *Growing computing power.* Training of state-of-the-art deep learning models is an intense task involving millions of parameters to optimize. GPUs are particularly well suited for the operations involved, and specialized software is available. See the supplemental material (Section S4) for further information.
- *Large datasets.* Part of the promise of deep learning is to take advantage of large datasets with millions of examples. An investigation into the effect of dataset size suggests a logarithmic relationship between performance on vision tasks and training data size [34].

When dealing with data that are known to have a certain structure (e.g., spatial structure, temporal structure), DNNs can be modified to create more specialized architectures that take advantage of said structures. In the following, we provide a brief overview of such specialized DNN architectures.

---

2. This excludes the input layer, which lacks learnable parameters.

3. This implies a many-to-many relationship between learned concepts and units representing them.

4. Due to the complexity of deep learning algorithms, it is difficult to explicitly determine their learning capacity [22, ch.5.2], but it can be thought of as being related to the number of parameters and layers [34].

2.1.2 Learning spatially: Convolutional neural networks

Sensor recordings of natural scenes often inherently contain a spatial structure along some dimensions. Convolutional neural networks (CNNs) introduce layers with specialized operations into DNNs, which take into account the spatial structure of the data to make the network more efficient.

The convolutional layer consists of several learnable kernels, which are convolved with the layer input to produce activations. In terms of the layer's units, this can be interpreted as *parameter sharing* between units, since the same kernel is applied over different spatial locations. The kernel's receptive field is typically much smaller than the input, which leads to *sparse connectivity* between units of adjacent layers [22, ch.9.2]. In contrast to the fully-connected equivalent, this means that convolutional layers only establish connections between units that are spatially close, which dramatically reduces the number of parameters. Since CNNs are inspired by the mammalian visual cortex [35], we primarily see 2D convolutions applied to image data. However, it is possible to apply convolutions along any dimension of the input data, including 1D (e.g., audio data) and 3D convolutions (e.g., video data).

After the convolution operation, a nonlinearity is applied, typically the rectified linear unit [31]. Pooling layers between convolutional layers facilitate nonlinear downsampling of layer activations, and make the network invariant to translations in the input [22, ch.9.3]. Invariances to other transformations such as rotation or scaling are not directly implied by the architecture, but can be learned in convolutional layers. Finally, it is common to add one or two fully-connected layers after several alternating convolutional and pooling layers (e.g., [36]).

Compared to fully-connected DNNs, training CNNs is less difficult due to the reduced number of parameters. They were trained successfully in the 1990s [37], while fully-connected DNNs were still believed to be too difficult to train [25]. Since a record-breaking performance [36] at the *ImageNet* Large Scale Visual Recognition Challenge (ILSVRC) [38] in 2012, CNNs have attracted much attention from researchers and the mainstream media.

2.1.3 Learning from sequences: Recurrent neural networks

When learning from sequential data (e.g., audio, video), the goal is to capture temporal dynamics in an efficient way that allows generalization to sequences of arbitrary length. This can be accomplished by sharing parameters across time, instead of re-learning them for every step. As mentioned previously, CNNs can accomplish parameter sharing in a shallow way, by applying the same kernel at different points in time. Recurrent neural networks (RNNs), on the other hand, introduce recurrent connections in time, allowing parameters to be shared in a deeper way [22, ch.10].

A basic RNN extends the feedforward architecture by allowing recurrent connections to exist within layers. Simply put, the previous model state can be regarded as an additional input at each temporal step, which allows the RNN to form a memory in its hidden state over information from all previous inputs [39]. RNNs have a representational advantage over hidden Markov models (HMMs), whose discrete hidden states limit their memory [40]. Even RNNs with a single hidden layer can be considered as very deep networks, which becomes clear when we imagine "unrolling" them along the time dimension. It turns out that this depth makes training RNNs considerably more difficult, as gradients tend to vanish or explode during training—especially when processing long sequences.

To deal with this problem, multiple specialized RNN architectures with gated units have been proposed. *Long short-term memory* (LSTM) RNNs [28] are successful at learning long-term dependencies by providing gate mechanisms to add and forget information selectively. *Gated recurrent units* (GRUs) are a gating mechanism proposed more recently in the context of sequence-to-sequence processing [41]. RNNs, especially LSTMs, have had a profound impact on how sequences of data are processed [24]. They are incorporated in state-of-the-art AI systems, for example in automatic speech recognition (ASR) [42].

2.1.4 Unsupervised learning models

To learn useful feature representations from data, the most common approach today is *supervised* learning: Researchers provide the learning algorithm with clues on how to improve parameters, typically in the form of corresponding data labels and a loss function measuring how "bad" a representation is. Taking a classification task for example, a useful representation would be one that makes the classes of interest linearly separable. However, labeling data is expensive and large unlabeled datasets are easier to come by. *Unsupervised* learning algorithms attempt to learn useful representations without being given explicit clues such as data labels—instead, a form of regularization is introduced.

An autoencoder (AE) [43] is a neural network that learns two functions, $f : x \mapsto z$ and $g : z \mapsto x$, to restore its input x from an intermediate representation z. The idea of the AE is to avoid identity mapping between f and g, either by forcing z to be of lower dimensionality than x in basic AEs, or by other forms of regularization, such as restoring x from a corrupted version $\tilde{x}$ of itself in denoising AEs. When multiple layers are between input or output and intermediate representation in an AE, we speak of stacked autoencoders (SAEs). Restricted Boltzmann machines (RBMs) [44] are undirected probabilistic models that learn a representation of their input in a layer of latent units. Deep Boltzmann machines (DBMs) [45] consist of several layers of latent units with undirected connections. A deep belief network (DBN) [23] on the other hand consists of several layers of directed connections, with an RBM as its final layer.

In practical applications, the lines between supervised and unsupervised learning are often blurred [22, p.105]. Unsupervised pre-training is a technique whereby single-layer unsupervised models, such as RBMs, are iteratively trained and stacked into a deep model [23]. Both supervised and unsupervised models can benefit from the learned hierarchy of feature representations: By adding a classification layer and supervised fine-tuning, it was found that difficulties in training fully-connected DNNs could be overcome [46]. This technique was responsible for the resurgence of deep learning since 2006, but has later gone out of fashion as it is no longer required for training fully-connected DNNs. Another use is found in initializing deep unsupervised models, such as DBNs and DBMs.

2.2 The notion of affect in affect recognition

The notions of affect and emotion[5] are subjective in nature. As a result, the question of how affective states should be represented is still an unsolved issue with no consensus reached in the literature. Two differing views are dominant: The *categorical* view of affect as discrete states, and the *dimensional* view of affect where states are represented in a continuous-valued space. Both categorical and dimensional models are also used to decode the nature of human emotion in specific brain regions (e.g., amygdala, insula), which is subject to an ongoing debate in affective neuroscience (see [48] for a review). Affective computing has largely stayed agnostic on the debate regarding the appropriateness of these views [9]. In practice, the type of affect labels available in databases is often chosen to most naturally fit the available sensor data—*categorical* if images or sequences are to be matched with a single affective state, and *dimensional* for continuous affect prediction (see Section 3.4).

The notion of categorical affective states originated in Charles Darwin's research on the evolution of affect, and remains much discussed in the literature [49]. Most categorical models assume the existence of *basic* affective states as building blocks of more complex ones [49]. In practice, Ekman's basic affective states derived from facial expressions [50] are most commonly used in the context of affect recognition; they comprise *anger*, *happiness*, *surprise*, *disgust*, *sadness*, and *fear*. Others include Plutchik's wheel of emotions [51], and the basic model [52]. In application-driven efforts, *non-basic* affective states are more commonly used, where specific user states and the intensity thereof are of interest—such as *boredom* or *frustration* in HCI [9], [11]. This approach sidesteps issues of theoretical validity to instead focus directly on the relevant non-basic affective states.

Dimensional models aim to avoid the restrictiveness of discrete states, and allow more flexible definition of affective states as points in a multi-dimensional space spanned by concepts such as affect intensity and positivity. This addresses the notion that discrete categories may not fully reflect the complexity of affective states and the associated problems in labeling and evaluating affective displays (e.g., labeler agreement) [11], [9], [53]. For affect recognition, the dimensional space is commonly operationalized as a regression task (e.g., for arousal [54], [18]) or as a classification task where the continuous space is discretized (e.g., for different levels of arousal, [55], [56]). Mappings can be established between dimensional and categorical models of affect (e.g., [57]). The most commonly used example is Russell's circumplex model [58], which consists of the two dimensions *valence* and *arousal*, sometimes extended by *dominance* and *likability*.

2.3 Frontiers in affect recognition

Research on affect recognition has seen considerable progress as the focus has shifted from the study of lab-based, acted databases to real-life scenarios since the 2000s [11]. In the 2010s, the focus has remained on such conditions, as research is exploring more complex models to better take

---

5. The terms *affect* and *emotion* are widely used synonymously in the context of affective computing [47].

advantage of available sensor data. The annual competitions AVEC and EmotiW have been established to focus on multimodal affect recognition, encouraging researchers to benchmark model accuracy in a fair manner. Multiple factors make building such models challenging:

- *Natural settings.* Data collected in uncontrolled real-life settings can introduce many additional sources of variation that need to be accounted for. Visually, these include occlusions, complex illumination conditions in different settings, spontaneous behavior and poses, as well as rigid movements. Audio may contain background noise, unclear speech, interruptions, and other artifacts.
- *Temporal dynamics.* The temporal dimension is an integral element of affective display that has not yet been fully embraced by research on affect recognition, especially FER. Temporal dynamics can be rich with contextual information that could be captured by models, e.g., to distinguish between displays that are similar in the short term, and to assess the relative importance of specific segments [7].
- *Multimodal sensor fusion.* Humans rely on multiple modalities when expressing and sensing affective states in social interactions. It seems natural that computers could benefit from the same variety of sensors [47]. In fact, there has been an increased interest to design such multimodal systems [59], and it is generally accepted that audiovisual sensor fusion can increase model robustness and accuracy. However, without knowledge of how exactly humans handle this problem, it is unclear how and at which level of abstraction the modalities need to be fused.
- *Limited availability of labeled data.* The increase of factors that make learning difficult—such as model size and the nature of ambitious affect recognition tasks—has outpaced the increase in availability of labeled examples. It is a challenging task to successfully train large DNNs with relatively few labeled data. Advanced regularization methods are necessary to avoid problems like overfitting. Transfer learning [60] attempts to transfer knowledge between emotion corpora and other domains such as object recognition. Furthermore, unsupervised and semi-supervised learning are being explored to access knowledge encoded in unlabeled datasets.

2.4 Towards learning deep models of affect

While traditional human affect recognition systems follow standard machine learning approaches, it seems intuitive that these algorithms may benefit from being more reminiscent of how the human brain works. Leading researchers in affective computing have supported this notion in the past, voicing the expectation that biologically inspired systems could be more suitable for affect recognition than human-engineered systems [1], [47].

In light of the challenges discussed in Section 2.3, current affect recognition tasks combine several characteristics of difficult perception problems in AI. The variation in the sensor data is dominated by factors mostly unrelated to the task, (i) spatially (e.g., in image data and short-term

audio segments), (ii) temporally (e.g., in image sequences and long-term audio), and (iii) distributed across modalities on different time scales. In order to perform the desired classification and regression tasks in affect recognition, it is necessary to obtain useful feature representations from the available multimodal high-dimensional sensor data. These feature representations should disentangle the underlying factors of variation in order to isolate the factors that distinguish affective displays [61].

Previous approaches rely on cleverly handcrafted features and shallow learning models to derive such feature representations. Unfortunately, the steps involved in designing handcrafted features can be work-intensive and error-prone. Motivated by the success of deep learning in other AI-relevant tasks based on images and speech [24], researchers have started to use deep models for representation learning in affect recognition. As discussed in Section 2.1, DNNs are more efficient at learning and representing the given functions, both statistically and regarding human involvement [20]. Labeled data and computation time are limited and costly, incentivizing the adoption of such models in affect recognition. Based on the previous discussion of challenges and model types, we identify three distinct ways in which deep learning is leveraged to learn useful feature representations for affect recognition tasks:

- *Learning spatial feature representations.* For images [61], short-term[6] image sequences [62] and audio segments [54], DNNs and especially CNNs are used to learn spatial feature representations.
- *Learning temporal feature representations.* To learn representations of the temporal dynamics found in audio [63], sequences of images [64], and physiological measurements [55], DNNs and especially RNNs are successfully applied.
- *Learning joint feature representations* for multimodal data. In multimodal approaches, DNNs are leveraged to learn joint feature representations from multiple unimodal feature representations to accomplish feature-level fusion [65].

After learning such feature representations, the derived features (or *deep features*) can be used as input for simple classification and regression methods, such as logistic regression and support vector regression (SVR). In Section 3, we give an in-depth breakdown of how deep learning models for human affect recognition are applied in practice.

3 THE STATE OF THE ART

Since around 2010, deep learning methods have started to fulfill crucial roles in human affect recognition systems. Indeed, it is safe to state that they are currently driving most state-of-the-art results in this field. Our goals are to (i) measure the adoption of deep learning in the field, and (ii) break down what specific functions are being fulfilled by DNNs in human affect recognition systems.

Because of these goals, we conducted a two-stage literature search[7] of studies on human affect recognition since

6. For this study, we consider segments of up to 1 s as short-term.
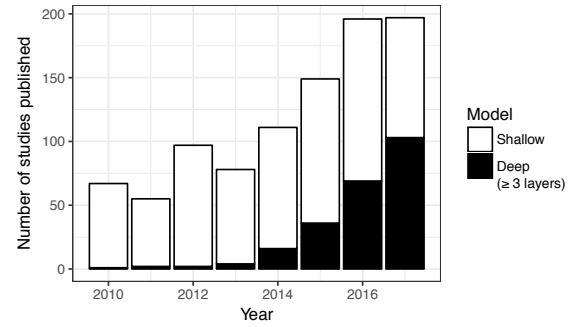7. Databases searched: The ACM Digital Library, IEEE Xplore, SpringerLink, and Web of Science.

Fig. 1. Stage 1: Number of studies on affect recognition published by year, distinguished by their use of deep or shallow learning models.

2010. This search is restricted to studies that use sensor data of cues directly given by the human body[8], i.e., facial expressions, movements, speech, and physiology. Further affect recognition fields that focus on affect communicated through means other than the human body—such as generic images and videos [66], music [67], [68], and text [69]—are out of the scope of this review. See [70], [7], and [11] for earlier reviews on human affect recognition.

Stage 1 of our search yielded a total of 950 studies. Following the terminology introduced in Section 2.1, these studies were then manually classified to indicate the use of shallow or deep models. Considering the temporal distribution of the studies, Fig. 1 illustrates two developments:

- Between 2010 and 2016[9], there was an overall increase in the number of studies on human affect recognition (25% year-on-year increase on average).
- Deep learning has gained considerable attention in this field since 2010: Up from one to two studies per year, it is being employed in 52% of studies in 2017—a 119% average year-on-year increase in the number of published studies.

In Stage 2, we focus exclusively on the 233 studies found in our review that use deep learning for affect recognition. They form the basis for the review in this section. These studies were further classified by (i) the usage of deep learning according to the three ways introduced in Section 2.4, and (ii) the modality used as a basis for recognizing affect. Table 1 lays out the result of these classifications by listing the numbers of studies falling into each category. Note that individual studies often use multiple modalities or apply deep learning in more than one way.

From Table 1, it is apparent that the application of DNNs for FER has attracted the most attention in the literature, featured in almost twice as many studies compared to SER, which is featured second most frequently. Considering physiological signals, DNNs are most frequently applied in studies based on electroencephalography (EEG).

In the central part of Table 1, we can see that learning feature representations of spatial information is the most

8. Since we only look at recognition of affective states, facial action unit detection and facial feature point detection are excluded. For SER, we restrict the scope to paralinguistic content, which is the main focus of research during the time frame covered by our review.
9. The search only contains partial data for 2017, hence 2017 is excluded here.

TABLE 1
Number of identified studies on human affect recognition by modality, and applications of deep learning.

Modality		Learning deep feature representations			Total[a]
		Spatial (Section 3.1)	Temporal[a] (Section 3.2)	Joint (Section 3.3)	
Visual	Facial expression	141	36	19	158
	Body movement	-	4	2	4
Auditory	Speech	35	64	17	82
Physiological	EEG	4	17	3	18
	Peripheral	-	6	4	7
	Other	-	2	3	3
Total[a]		173	105	21	233

[a] Totals may not equal row/column sums due to overlaps between modalities and approaches.

common application of DNNs in human affect recognition to date, especially in FER. For SER in particular, learning of temporal feature representations with DNNs is an active research area. A less active, but developing application area of DNNs in human affect recognition is learning joint feature representations for the purpose of early *feature fusion* across different modalities.

3.1 Learning spatial feature representations

The goal in spatial feature learning is to learn expressive feature representations of data with spatial structure. In practice, we find that deep architectures are frequently applied to exploit this characteristic in sensor recordings containing static and short-term cues of affective behavior—both visual imagery and short segments of audio and physiological data can be interpreted in this way. Some approaches combine handcrafted features with deep architectures such as fully-connected DNNs (see Fig. 2, S1a–S1b). As discussed in Section 2.1.2, the design of CNNs is based around the prior of spatial coherence. Hence, CNNs are the most popular approach for learning spatial features (see Fig. 2, S2a–S2e).

3.1.1 Learning spatial features for FER (S1a, S2a, S2b)

Mehrabian [71] famously posited that 55% of the emotion conveyed in a message is perceived visually. Indeed, FER is most prominently featured in our search results on spatial feature learning, as Table 1 reveals. Detailed surveys are available [72], [73], giving an overview of the field in general. In this section, we discuss the application of deep learning across 141 studies to learn spatial features from images. We distinguish between approaches using fully-connected DNNs (Fig. 2, S1a), and CNNs (Fig. 2, S2a–S2b).

Conventional approaches rely on handcrafted features to represent faces by their *shape* or *appearance*. Shape representations use explicit knowledge about facial geometry to encode a given expression, such as the location of certain facial feature points. The most common appearance features, such as local binary patterns (LBP), local phase quantization (LPQ), and histogram of oriented gradients (HoG), encode low-level texture information in local histograms. Other methods of feature extraction include convolving the input with handcrafted Gabor filters and scale-invariant feature transform (SIFT). As Sariyanidi et al. [72] pointed out, such handcrafted features focus on low-level description of edge
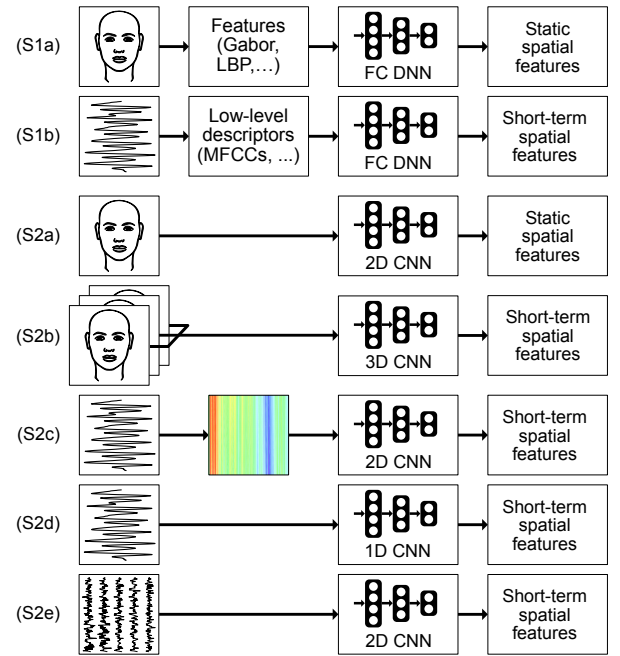

Fig. 2. Applications of deep learning for spatial feature learning with fully-connected DNNs (S1a–S1b) and CNNs (S2a–S2e).

distributions. While they provide robustness against illumination variations, they are less suitable for discrimination between high-level concepts such as facial features. On the contrary, CNNs natively learn a hierarchy of features that builds from low-level to high-level representations. While the first layer learns general concepts similar to Gabor filters, the last layers learn more specific concepts that tend to be semantically interpretable. As a result, 93% of studies reporting direct comparisons find that deep spatial features outperform handcrafted spatial features for FER.

Learning spatial features from intermediate handcrafted features (S1a in Fig. 2). As evident from Table 2, especially during the initial adoption of DNNs in FER, handcrafted features and DNNs were combined in a sequential way. In the first step, this approach extracts low-level handcrafted features from pixel values. Appearance features such as LBP (e.g., [74], [75]) and Gabor features (e.g., [76], [77]) are preferred for this approach. Due to the reduced dimensionality, it is then feasible to apply fully-connected DBNs (e.g., [76])

TABLE 2
Learning spatial feature representations: Number of studies adopting
different approaches over time.

Year	FC DNN		1D/2D/3D CNN				
	S1a	S1b	S2a	S2b	S2c	S2d	S2e
2012	-	-	1	-	-	-	-
2013	-	1	1	-	-	-	-
2014	3	1	4	-	1	-	-
2015	3	-	24	2	2	-	-
2016	2	2	42	-	5	3	3
2017	3	2	57	9	11	9	1
Total	11	6	129	11	19	12	4

Fig. 3. Number of hidden layers in CNNs for spatial feature learning in FER.

and SAEs (e.g., [77], [78]) for unsupervised learning of high-level features.

Learning spatial features directly from 2D image with CNNs (S2a in Fig. 2). CNNs are well suited to learn spatial features directly from image pixels. As becomes clear from Tables 1 and 2, this approach dominates our search results. Due to a lack of understanding why deep learning works well in practice [79], and limited availability of labeled data, the challenge faced by researchers is choosing an appropriate architecture. Fig. 3 provides an overview of typical choices regarding the number of fully-connected and convolutional layers, and indicates whether transfer learning was used. We can distinguish between two approaches: Region I in the left of Fig. 3 refers to architectures totalling six or less convolutional and fully-connected layers. These CNN architectures, which are specifically designed for FER, make up 56% of the studies. Most of these studies rely on smaller model sizes instead of transfer learning to avoid overfitting the relatively small number of examples.

Region II consists of architectures with more convolutional layers, and hence potential to learn more expressive features. Many of these are existing architectures that have proven successful for other tasks such as object recognition. Most frequently chosen are VGG Net [80] (23 studies) and AlexNet [36] (18 studies). AlexNet is the architecture that won the 2012 ILSVRC [38]. It consists of five convolutional and three fully-connected layers, with a total of 60 million parameters. VGG Net, an entry at the 2014 ILSVRC, is a deeper CNN coming in variants of 16 and 19 layers. Other choices include GoogLeNet [81] (winner of ILSVRC 2014), ResNet [29] (winner of ILSVRC 2015), and Tang's winning entry at ICML 2013's FER challenge [82]. Most of these studies use additional datasets and transfer learning to avoid overfitting.

Learning short-term spatial features from image sequences with 3D CNNs (S2b in Fig. 2). Building on the concept of spatial representation for a single 2D image, this approach interprets an image sequence as a spatio-temporal volume. Standard CNNs can be extended to accept 3D volumes as input by increasing filter dimensionality to support spatio-temporal convolutions. Such *3D CNN* architectures [83] are theoretically capable of learning spatio-temporal features such as motion of facial action units. Since the number of model parameters and therefore the number of required examples increase with temporal depth of input sequences,
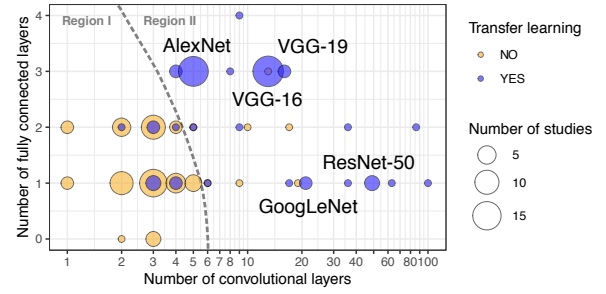
the chosen number of frames is typically quite low[10], and 3D CNNs are limited to short-term sequences (up to 1 s). The nature of this approach also requires that input sequences consist of a standardized number of frames. This means that researchers need to downsample or interpolate videos [62].

Group-level FER. Group-level affect recognition is a sub-discipline of FER, where the goal is to assess the overall expression of all persons in an image. It has been featured in the EmotiW competition in 2016 [12] and 2017 [86]. For this purpose, spatial features are typically extracted for each person, and fused in some way—e.g., by considering multiple faces as a sequence and applying an LSTM [15].

Complementarity of deep and handcrafted spatial features. In benchmarking experiments across various datasets, many studies found that features extracted with CNNs lead to higher recognition accuracies than handcrafted features (e.g., [87], [88], [89], [90], [91], [64], [92], [93], [94]). However, several challenge-winning studies [95], [15], [17], choose to use both handcrafted and deep features, e.g., by score-level [95] or model-level [17] fusion of separate handcrafted and deep models. This suggests that deep and handcrafted features are complementary. As of this writing, the findings of most studies doing related comparisons support this assumption [96], [97], [98], [99], though in many cases a comparison is difficult as reported accuracies for well-performing fusion approaches include features from other modalities (see Section 3.3). Only one study reported that deep and handcrafted features are not complementary [100].

Pre-processing to simplify the FER learning task. Instead of learning from unprocessed images, the majority of the reviewed studies apply some form of pre-processing to image data. These steps reduce the amount of variation the model has to account for, and thus simplify the learning task. *Face cropping* is standard practice, and reported to increase model accuracy (e.g., from 54% to 72% in [94]; see also [15]). This involves the detection of the face and feature points, although some databases come with pre-detected or cropped faces (e.g., [101], [102], [103]). *Spatial normalization* techniques include face alignment and face frontalization: Simple adjustment of face rotation and facial feature point alignment is reported to improve model accuracy (e.g., from 54% to 62% in [94]; see also [15], [99]). More advanced face frontalization involving the approximation of 3D shape is useful when dealing with 3D head pose variation [104], [105]. *Intensity normalization*, on the other hand, aims to

---

10. The reviewed studies used between 3 [84] and 16 [85] frames.

normalize illumination related factors such as brightness and contrast, which can be an issue with images taken in the wild or across multiple databases. Researchers report improvements in accuracy by applying intensity normalization (e.g., from 54% to 57% in [94]; see also [98]).

Limited availability of labeled data. Since the number of labeled examples for FER remains relatively limited, large[11] models are likely to overfit the data [106], [107], [89], [108]. Several techniques are available to address this problem: (i) transfer learning, (ii) data augmentation, and (iii) architecture- and training choices promoting regularization, such as dropout. In *transfer learning*, the goal is to use additional corpora to learn generic visual descriptors that are found to be effective in improving initial model parameters [109], and thus reduce overfitting. Especially when adopting large models originally intended for object recognition, it is common to pre-train on a large-scale database such as *ImageNet* [110] (14M annotated images). Another database frequently chosen for pre-training is *VGG-Face* [111] (2.6M face images). Smaller, more relevant databases such as *FER2013* [103] and *CK+* [112] are also frequently used for pre-training. To make use of generic visual features, authors acquire large models pre-trained for object classification, "freeze" the lower-level layers, and fine-tune a selection of higher-level layers for affect recognition. Improvements in accuracy are reported when following this approach (e.g., from 39% to 42% in [89]; see also [99], [92]).

Data augmentation is a technique whereby existing images or sequences are manipulated to reduce overfitting. Researchers either use static rules to generate new examples for the same original label, or manipulate images randomly before training. Such manipulations include horizontal flipping, cropping, rotation, translations, changes to color, brightness, and saturation, as well as scaling. This way, researchers artificially increase the number of available examples or training epochs by a factor typically between 10 and 30 [88], [113], [114], and up to 300 [115]. Studies running experiments on this technique report accuracy improvements (e.g., from 79% to 89% in [87]; see also [74], [94]). *Dropout* [116] is a technique that reduces overfitting by randomly dropping out neurons during training, thus forcing the network to learn redundantly. It is widely used in the reviewed studies—57% report using dropout for fully-connected layers, and 12% report using it for convolutional layers. Khorrami et al. [87] reported an increase in accuracy of 2.5% after applying dropout to fully-connected layers.

3.1.2   Learning spatial features for SER (S1b, S2c, S2d)

Beyond spoken words, the acoustic properties of human speech are rich with information about the speaker, such as gender, age, and affect (see [117], [118] for comprehensive reviews). In this section, we focus on the 35 studies that employ deep learning for spatial feature learning in SER. The potential of replacing or complementing traditional short-term descriptors with DNNs in speech related classification tasks was pointed out as early as 2009 [119]. Especially CNNs are found useful in modeling speech features [120].

We distinguish between approaches using fully-connected DNNs (see Fig. 2, S1b) and CNNs (see Fig. 2, S2c–S2d).

Research has shown that short-term spectral, prosodic, and energy features of speech carry affective information [121]. In conventional SER approaches, it is common practice to capture such properties using handcrafted features known as *low-level descriptors* (LLDs). LLDs are sampled from small overlapping audio segments or *frames*; a common choice is a window size of 25 ms and a step size of 10 ms [122]. Most recent models use pre-defined sets of LLDs for spatial modeling. Standard sets such as eGeMAPS [121] and ComParE [123] typically include cepstral descriptors such as *Mel-frequency cepstral coefficients* (MFCCs), energy-related descriptors such as *shimmer* and *loudness*, frequency-related descriptors such as *pitch* and *jitter*, and spectral parameters. They can be extracted with software tools such as openSMILE [124] and openEAR [122]. This review highlights how DNNs are used in the state of the art to complement and replace handcrafted LLDs. Overall, 90% of studies reporting direct comparisons find that deep spatial features outperform handcrafted spatial features for SER.

Learning spatial features from handcrafted feature representations of speech (S1b in Fig. 2). A limited number of early applications in SER were combinations of DNNs with handcrafted LLDs. The idea was to use fully-connected DNNs to replace Gaussian Mixture Models (GMMs), which occupied the role of short-term modeling in the commonly used GMM-HMM architecture from ASR. For example, Li et al. [125] used a six-layer DNN to learn frame-level features from concatenated MFCCs of a sliding context window. In their experiments, this yielded an accuracy improvement of more than 10% over GMMs.

Learning spatial features from raw spectral representations of speech with CNNs (S2c in Fig. 2). Many recent studies in SER leverage DNNs to avoid the step of handcrafted feature engineering (see Table 2). Mirsamadi et al. [63] pointed out that most commonly used frame-level LLDs in SER can be derived from spectral representations of the raw speech signal. Without any feature engineering, they were able to learn features similar to LLDs from the raw spectral representation of individual audio frames at 25 ms, leading to an accuracy increase of 4%. Such learned features can be shown to have similarities to handcrafted LLDs [54].

Spectrogram representations are computed using multiple frames and allow speech segments to be interpreted as 2D images. They can be based on Fourier transform of the raw waveform (e.g., [126], [127]) or minimally hand-engineered on the log Mel-frequency cepstrum representation (e.g., [85], [128]), which closer matches the characteristics of human auditory perception. CNNs can be applied to directly learn features from such representations, which are typically between 250 ms and 1 s in length. Since the suggested minimum time required to identify affect from speech is quoted in the literature as 250 ms [129], [85], the resulting features can directly be used for classification of short utterances (e.g., [126], [130]), or be regarded as short-term features (e.g., [85], [131]) for further temporal modeling as discussed in Section 3.2.3.

Most studies chose custom architectures of one to three convolutional layers and one to three fully-connected layers for this task. Some considered using known architectures

---

11. When using CNNs for FER, researchers tend to resort to methods like transfer learning in Region II, see Fig. 3. Note however that the number of trainable parameters can be subject to many other factors.

from object recognition [132], [85], [127]. Here, the authors regarded it as necessary to pre-train these larger models on the ImageNet database to avoid overfitting the relatively few labeled examples.

Learning spatial features from the raw waveform with CNNs (S2d in Fig. 2). Feature learning directly from the raw waveform was proposed in 2011 [133]. Since 2016 (see Table 2), a number of studies have started applying this idea in SER. CNNs can be applied to raw 1D audio (see Section 2.1.2); indeed, all except one study in our review use CNNs for this task. Trigeorgis et al. [54] were the first to do so: They used two convolutional layers for spatial modeling, almost doubling ground truth correlation over LLDs for arousal, and slightly improving for valence. Bertero et al. [134] proposed one convolutional layer with 25 ms kernels, which achieved a 3% accuracy improvement over LLDs.

Limited availability of labeled data. To avoid overfitting, both dropout [116] and batch normalization [33] can help to achieve better regularization. Dropout is applied in fully-connected (reported in 43% of studies, e.g., [135]) and convolutional layers (14% of studies, e.g., [108]). Multiple studies have shown that transfer learning can improve model accuracy by leveraging additional sources of related knowledge (e.g., from other paralinguistic tasks [136], various standard databases [137], and different affect representations [135]). SoundNet [138], a 1D CNN trained with unlabeled video, has been shown to perform well in SER even without fine-tuning [139], and was featured in a challenge-winning submission [17]. Semi-supervised learning can give access to knowledge contained in unlabeled datasets [140]. Knowledge transfer from domains like music [141] and visual object recognition [132], [85] is also possible. Data augmentation to artificially increase the dataset size is used less often than for FER; notable examples include the addition of Gaussian noise [129], different sampling frequencies [130], and modified playback speed [142].

3.1.3 Learning spatial features from physiology (S2e)

As early as 2001, physiological responses were shown to convey information about affective states in machine learning [8]. Candidates include measures of the peripheral physiology via electrocardiography (ECG), electrodermal activity (EDA), and brain activity via EEG. However, affective computing research initially focused mostly on FER and SER, partially due to a lack of interest and inconvenient sensors [7]. More recently, interest in physiological affect recognition is seeing a resurgence [143], owed in part to the capabilities of modern, portable monitoring devices [144].

As part of our review, we identified only four studies using 2D CNNs to learn spatial features from EEG data (see S2e in Fig. 2). All achieved accuracy improvements over handcrafted approaches, but a lack of training data was also mentioned [145], [146]. For example, Yanagimoto and Sugimoto [145] divided the raw 16-channel EEG data into 1s segments and used a seven-layer CNN with 10 ms kernels on the first layer, leading to accuracy improvements of over 20%. Similar to some work in SER, Li et al. [55] considered a spectrogram representation of the EEG signal at a frame size of 1s. Another way to learn spatial features from the EEG signal is to reflect it as a 2D map representing the location of the electrodes on the head [147].
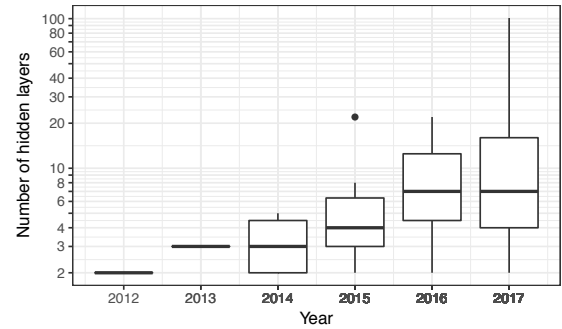

Fig. 4. Number of hidden layers used for spatial feature learning in FER.

3.1.4 Takeaways for spatial feature learning

- Deep spatial features lead to higher accuracies than handcrafted spatial features. Out of 103 studies that reported comparisons, 93% support this finding.
- However, in contrast to fields such as object- and speech recognition, both are often found to be complementary in affect recognition as of this writing. This suggests that the full potential of deep learning in affect recognition may not have been seen yet.
- CNNs are the most widely used architecture for spatial feature learning (91% of the studies in our review; see Table 2). Instead of using spectrogram representations, recent research starts to apply CNNs directly to raw speech and physiological data.
- To achieve higher accuracy, research strives towards "deeper" models (Fig. 4 illustrates this for FER), but runs into the problem of overfitting. This is the main challenge for current research.

3.2 Learning temporal feature representations

When learning from sequences, the goal is to learn feature representations that capture temporal dynamics [40]. This allows models to consider the temporal variation of spatial characteristics in sensor data (e.g., in SER [118] and video-based FER [106]). As discussed in Section 2.1, both CNNs and RNNs provide architectures that can learn representations of sequences of data. We found that the existing architectures—spanning all studies and modalities—can be classified into one of three approaches illustrated in Fig. 5: (T1) fully-connected DNNs for learning spatio-temporal features from aggregated frame-level spatial features, (T2) RNNs for global temporal modeling based on frame-level spatial features, and (T3) CNNs for local temporal modeling.

3.2.1 Learning temporal features for FER

A straightforward approach to derive sequence-level features from video data is to first extract high-level spatial features (such as facial characteristics, see Fig. 7) from individual frames, and then aggregate these in some way. This can be achieved by simple feature pooling strategies such as mean pooling, max pooling, or feature concatenation. However, such strategies typically ignore most of the temporal variation in the sequence, which may contain valuable contextual information. Well-designed models seek
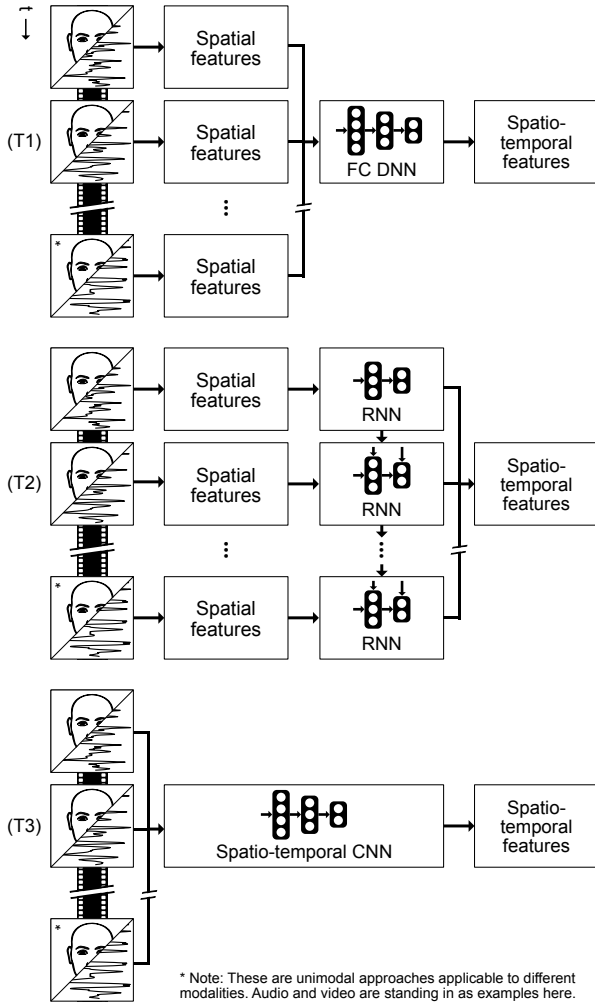
Fig. 5. Applications of deep learning for temporal feature learning with fully-connected DNNs (T1), RNNs (T2), and CNNs (T3).

TABLE 3
Learning temporal feature representations: Number of studies adopting different approaches over time.

Year	FC DNN (T1)[a]			RNN (T2)[a]			CNN (T3)[a]		
	V	A	P	V	A	P	V	A	P
2010	-	-	-	-	1	-	-	-	-
2011	-	1	-	1	1	-	-	-	-
2012	-	-	-	1	1	-	-	-	-
2013	-	-	-	-	-	-	-	-	1
2014	-	2	2	2	3	1	-	-	-
2015	2	3	2	1	-	-	1	1	-
2016	3	10	4	10	5	2	-	5	2
2017	1	11	6	15	13	1	3	12	1
Total	6	27	14	30	24	4	4	18	4

[a] Modalities: V = Visual (FER, Body), A = Audio (SER), P = Physio (EEG, Peripheral, Other).

to further exploit such information. To some extent, this is possible with common handcrafted features: Appearance features can be extended for spatio-temporal representation by considering a third orthogonal plane [96], [99]. With DNNs, temporal modeling capabilities in FER can further

be improved: 94% of studies reporting comparisons with handcrafted approaches find that deep temporal features perform better for FER.

Learning spatio-temporal features from aggregated frame-level spatial features (T1 in Fig. 5). In some cases, fully-connected DNNs are applied to achieve dimensionality reduction on high-dimensional spaces of aggregated handcrafted features. For example, Zhang et al. [148] and Ranganathan et al. [149] used fully-connected DBN models to learn from aggregated facial feature point trajectories, both improving recognition accuracies over shallow aggregation strategies.

Global temporal modeling with RNN based on frame-level spatial features (T2 in Fig. 5). The properties of RNNs, as discussed in Section 2.1, make them well-suited to model the temporal variation of frame-level spatial features. This approach first extracts high-level spatial features from each face image, which are then considered as sequential input to the RNN. Advantages of this approach include the ability to process long sequences, and the possibility of both sequence-level and continuous frame-level affect recognition on image sequences of arbitrary length. Early on, RNNs were used to learn temporal context from handcrafted spatial features such as coordinates of facial feature points [150], optical flow [129], and LBP [151].

More recent studies combine RNNs with deep methods for spatial feature learning discussed in Section 3.1.1, by adopting deep features from the last layer of a CNN trained for affect recognition (e.g., [106], [18], [64]). We see both CNN-RNN (e.g., [18], [90]) and CNN-LSTM (e.g., [108], [105], [64]) architectures, with CNN-LSTM being the more frequent choice among the reviewed studies. Global temporal modeling is found to lead to improved accuracies when compared with simpler methods such as pooling of spatial features (e.g., [152], [90], [108]). A disadvantage of most CNN-LSTM implementations is that training occurs in a disconnected way: The CNN is trained on frame-level, specifically for static spatial affect recognition. Hence, the extracted features are not necessarily optimal for further temporal context learning by the RNN. End-to-end training of the entire CNN-LSTM system addresses this problem, and can lead to accuracy improvements [108].

Local temporal modeling with 3D CNN (T3 in Fig. 5). When using 3D CNN for spatio-temporal modeling of image sequences as discussed in Section 3.1.1, the line between spatial and temporal representation learning can be blurred. While this approach is typically limited to very short sequences, with further pooling steps necessary to derive sequence-level labels (e.g., [84], [85]), in some cases spatio-temporal features can be derived for entire (short) sequences. For example, Gupta et al. [62] used a variant called *slow fusion* [153], which treats the time domain like a spatial domain, progressively learning low-level to high-level temporal features. As the amount of parameters required due to the temporal depth of the input is effectively reduced, this allows for more input frames.

3.2.2 Learning temporal features from body movement

Besides facial expression, body movement and gestures are other means of expressing affect visually [1]. In the reviewed studies, spatial features representing such movements are extracted using skeletal and shoulder tracking. For example,

Ranganathan et al. [149] and Kaza et al. [154] used the approach illustrated in Fig. 5 (T1), to learn spatio-temporal features from statistics of skeletal tracking point trajectories. Shoulder cues were used by Nicolaou et al. [150] in the RNN approach illustrated in Fig. 5 (T2). In comparison with facial expression and speech, they were found to be less expressive for prediction of both arousal and valence.

3.2.3 Learning temporal features for SER

To derive fixed-length features at the utterance level, SER models traditionally aggregate LLDs by *high-level statistical functionals* (HSFs) such as mean and standard deviation. Standard sets of HSFs and LLDs are given in eGeMAPS [121] and ComParE [123]. HMMs have long served as a standard choice for further modeling of temporal variation in speech signals, especially in ASR [155], but also in SER [125]. More recently, RNNs have emerged as a preferred way of modeling the sequential aspect of speech [156]. In particular, 91% of studies reporting comparisons with handcrafted approaches find that deep temporal features perform better for SER. This review highlights how DNNs can complement or replace both HSFs and HMMs for learning temporal representations in SER.

Learning from aggregated frame-level features (T1 in Fig. 5). Since there is no consensus in the literature over a "universal" handcrafted feature set with superior performance [121], many recent studies have applied a "brute-force" approach, resulting in a large number of features per utterance. This number varies from several hundred [157] to several thousand [158], [123], depending on the employed LLDs and HSFs. DNNs can be integrated to learn more high-level representations of these handcrafted spatio-temporal feature spaces. Studies aiming to reduce feature space dimensionality with deep learning almost exclusively use fully-connected DNNs, consisting of two to four hidden layers. It is common to initialize model parameters layer-wise via unsupervised pre-training as RBMs (e.g., [159], [160]), or AEs (e.g., [78]); subsequently, a Softmax classification layer is added for supervised fine-tuning. Alternatively, DBNs or SAEs can serve as feature extractors for classification via support vector machine (e.g., [137]). Dimensionality reduction of handcrafted features with DNNs can lead to improvements in accuracy over various databases [161].

Global temporal modeling based on frame-level features with RNNs (T2 in Fig. 5). In this approach, an utterance-level RNN models the temporal variation of frame-level features. Most straightforwardly, LLDs can directly be fed into the RNN at the frame level (e.g., [141], [162], [63]). Depending on the nature of the source audio, it can be beneficial to apply HSFs to frame-level LLDs according to a sliding window before applying the RNN [151], [65]. One study suggested that a smaller window size (2s) could be the best choice [65]. In general, the addition of RNN for temporal modeling is associated with an increase in model accuracy (e.g., [129], [97], [162]). When dealing with dimensional labels, this allows learning features at the frame level. Here, LSTM is found to outperform state-of-the-art techniques like SVR (e.g., [150], [151]).

Since 2016, eight studies have explored combining deep spatial features and RNN-based temporal feature learning. For spectrogram-based spatial features, Lim et al. [163] found that the CNN-LSTM architecture yields the best result on Emo-DB [164]. Applications of similar architectures based on the raw waveform have also been attempted [54], [108], showing that end-to-end learning can outperform shallow models. Overfitting is still a problem for this approach due to the large number of model parameters and limited dataset sizes. Mirsamadi et al. [63] found that model performance is slightly lower with joint learning of both short-term spatial features and temporal context on the IEMOCAP dataset [165], while both improve model performance when applied independently.

Local temporal modeling with CNNs (T3 in Fig. 5). When CNNs are used for modeling speech, they typically combine spatial modeling in the short term with temporal modeling of longer segments or entire utterances. The kernels of higher-level (i.e., second or third) convolutional layers can often be interpreted as learning temporal structure based on spatial features learned by the kernels in the first layer (see Section 3.1.2). For example, Trigeorgis et al. [54] performed pooling across time after learning spatial characteristics from the raw signal in the first layer, and added a second layer with 500 ms kernels to learn temporal characteristics. Similarly, Zhang et al. [85] used an AlexNet to add increasingly more temporal context to learned feature representations. It is worth noting that while some studies directly used CNN features for affect prediction [142], [128], others combined local temporal modeling with global temporal modeling via RNN [108], [163], or pooling approaches [85].

3.2.4 Learning temporal features from physiological data

Learning from handcrafted features with fully-connected DNNs (T1 in Fig. 5). As highlighted in Table 3, approach T1 is the primary application of DNNs for feature learning from physiological data. For this purpose, fully-connected DNNs are initialized by iterative training and stacking of unsupervised models such as RBMs or AEs [166], and applied to functionals of frame-level spatial features. Typical for EEG are handcrafted features derived from the frequency domain, such as power spectral density (PSD) coefficients of different frequency bands. Zheng et al. [167] found that DBNs can improve recognition accuracy of models based on differential entropy features. Similarly, Xu and Plataniotis [168] showed that DBNs can build on PSD features to outperform state-of-the-art methods on the DEAP dataset. Deep learning has also been used as part of ensemble methods [169], and in the form of Echo State Networks [170] for dimensionality reduction of handcrafted EEG features. Yin et al. [171] used stacked autoencoders (SAEs) to learn high-level representations from various peripheral sensors including skin temperature and blood volume pressure, improving the state-of-the-art by 5%.

Learning temporal context from spatial features with RNNs (T2 in Fig. 5). RNNs have been used to learn temporal context from EEG features to improve recognition accuracies [146], [55]. Brady et al. [18] found that learning temporal context with LSTM and handcrafted features leads to improvements over shallow baseline models. Ringeval et al. [65] used a similar approach. They found that while the given physiological signal has lower predictive power than audiovisual signals, both are complementary.

TABLE 4
Learning joint feature representations: Number of studies adopting different approaches over time.

Year	FC DNN (J1)[a]				RNN (J2)[a]			
	VA	VAP	VP	PP	VA	VAP	VP	PP
2011	-	-	-	-	1	-	-	-
2012	-	-	-	-	1	-	-	-
2013	1	-	-	-	-	-	-	-
2014	-	-	-	-	-	1	-	-
2015	2	1	1	-	-	-	-	-
2016	2	1	-	1	1	-	-	-
2017	3	-	-	1	4	-	-	-
Total	8	2	1	2	7	1	0	0

[a] Modalities: V = Visual (FER, Body), A = Audio (SER), P = Physio (EEG, Peripheral, Other).

Learning spatio-temporal representations from raw data with CNNs (T3 in Fig. 5). A limited number of four studies have attempted to learn spatio-temporal features directly from raw physiological data for discrimination between affective states. Yanagimoto and Sugimoto [145] used a CNN on raw 16-channel EEG data to differentiate between positive and negative affective states, which is shown to outperform shallow models based on common features. Similar results are reported when learning from intermediate representations based on differential entropy [55]. Martinez et al. [172] were the first to learn deep features directly from the peripheral physiology. For this purpose, they used CNNs trained in an unsupervised way via AEs to learn features from raw blood volume pulse and skin conductance signals, which outperformed models based on handcrafted features.

3.2.5 Takeaways for temporal feature learning

- Deep temporal features lead to higher accuracies than handcrafted temporal features. Out of 73 studies that reported comparisons, 92% support this finding.
- While CNNs are well suited for local temporal modeling, RNNs are found to be useful for global temporal modeling of affect.
- Since 2015, there are studies using deep learning for both spatial and temporal feature learning.
- We are starting to see studies implementing end-to-end training for such models [54], [108], however in this setting the problem of limited labeled data becomes especially noticeable [63], [139].

3.3 Learning joint feature representations

It is generally accepted in the literature that multimodal (e.g., audiovisual) sensor combinations have complementary effects and thus may increase model accuracy [11]. The challenge in joint multimodal feature learning is how and at what stage to fuse data from multiple modalities. This challenge is complicated by the high dimensionality of raw data, differing temporal resolutions, and differing temporal dynamics across modalities. Surveys on the general problem of sensor fusion [173] and specifically on fusion for affect recognition [59], [174] are available.

Fusion can be achieved at early model stages close to the raw sensor data, or at a later stage by combining independent models. In early or *feature-level* fusion, features are
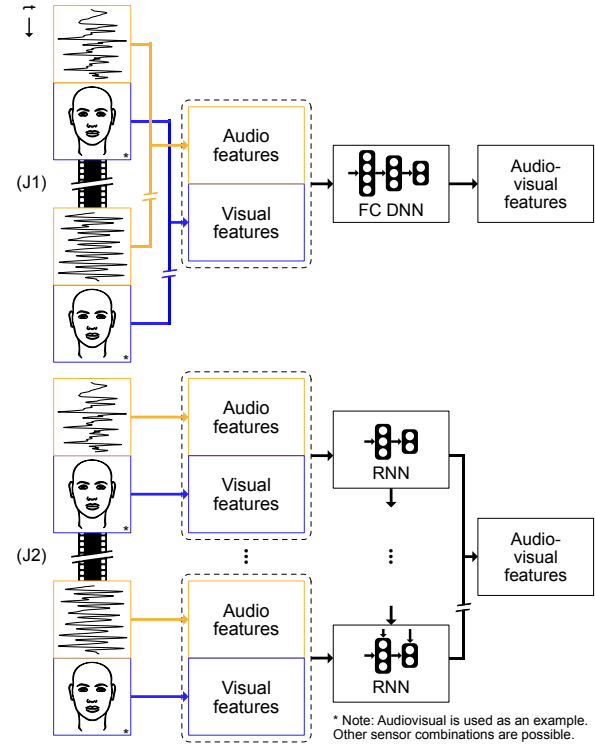

Fig. 6. Applications of deep learning for joint multimodal feature learning with fully-connected fusion DNNs (J1) and fusion RNNs (J2).

extracted independently and then concatenated for further learning of a joint feature representation; this allows the model to capture correlations between the modalities. Late or *decision-level* fusion aggregates the results of independent recognition models. To date, the literature generally reports that decision-level fusion works better for affect recognition given the datasets and models currently used [65]. While decision-level fusion typically only involves simple score weighing, feature-level fusion is a representation learning task that may benefit from deep learning. Here, we report on the approaches of 21 studies that use deep learning for joint feature learning from multimodal data.

3.3.1 Learning joint features with audiovisual data

The most common sensor combination found in 18 studies involves facial expressions and speech.

Feature-level fusion with fully-connected DNNs (J1 in Fig. 6). In this approach, joint feature representations are learned without considering the temporal context for fusion. For both modalities, video-level features are extracted using FER and SER methods that may involve both handcrafted and deep features (see Sections 3.1 and 3.2). A fully-connected DNN, typically initialized via unsupervised pre-training, then learns a high-level joint feature representation of both modalities as an improvement over "shallow" feature fusion. Kim et al. [159] and others (e.g., [149], [148]) demonstrated how this can be achieved with DBNs. This approach is feasible especially in cases where the goal is to label each video with one affective state. Alternatively, joint feature representations can be learned at the frame level, and then aggregated to the video level: Zhang et al. [85]

used a DBN to fuse frame-level audiovisual features learned independently via CNNs; the learned features are average-pooled for classification at the video level and lead to an improvement over state-of-the-art methods.

Feature-level fusion with RNNs (J2 in Fig. 6). Especially when predictions are required at the frame level for dimensional affective states, feature-level fusion could benefit by taking into account the temporal context. Modeling via RNNs makes this possible, potentially improving model robustness and helping to deal with temporal lags between modalities [162]. Initial studies reported that dynamic feature fusion can lead to performance improvements compared to simpler fusion strategies [162]. However, several other studies based on handcrafted features found that decision-level fusion on top of individual LSTM models leads to better performance [65], [129]. Learning from raw audiovisual data with two CNNs, Tzirakis et al. [108] used a two-layer LSTM network for feature fusion, which was found to outperform the state of the art.

3.3.2 Learning joint features with physiological data

A small number of three studies combined the audiovisual and physiological (AVP) modalities. Ranganathan et al. [149] demonstrated the feasibility of learning joint feature representations of AVP sensor data with approach J1 and a DBN, but do not compare the performances of different modality combinations. Ringeval et al. [65] used approach J2 with the AVP modalities and LSTM. They concluded that in feature-level fusion, ECG data helps for prediction of valence, but not arousal.

Feature-level fusion can also be based solely on physiological measurements. Yin et al. [171] successfully used a fusion SAE to aggregate handcrafted features from several different sensors. Similarly, Liu et al. [175] used handcrafted features derived from EEG and eye tracking as input into a SAE. Both studies found that the representation learned through feature-level fusion leads to improved accuracy over individual modalities.

3.3.3 Takeaways for joint feature learning

- Joint feature learning is most commonly applied to audiovisual fusion (see Table 4).
- To date, there is no consensus whether feature-level fusion with deep learning leads to superior accuracy over simple decision fusion. Out of 16 studies that reported comparisons, only 69% find that it does.
- While dimensional models of affect are only used in 10% of spatial and 30% of temporal feature learning studies, they are employed in 52% of studies on joint feature learning.

3.4 Databases and competitions

Most researchers rely on publicly available databases of affective display as source material for their studies. Of the 233 studies in our review, only 11% involved private databases not available to the public. A total of 77 different public databases were used across the reviewed literature. The specifics of these databases have considerable impact on algorithm design for affect recognition, which is why a comprehensive overview of databases and their properties

is essential to further understanding of the field. The main differences lie in the available modalities, the number of subjects and examples, details on how data was acquired, how affect was elicited and annotated, as well as the type of affective states used for labels. Some databases are more frequently mentioned due to them being featured in competitions. In Table 5, we give a summary of the 15 most commonly used databases in the reviewed studies.

As expected from previous findings, the visual modality is featured most frequently. Some databases focus exclusively on static FER with discrete labels of categorical affective states. Here, more recent databases such as FER2013 and SFEW2 tend to contain more examples than older databases such as JAFFE and CK+—this is made possible by resorting to sources like the web and semi-automatic labeling procedures as opposed to manual annotation of data collected in a laboratory. Audiovisual databases are a second type evident from the literature. A typical setup for earlier instances (e.g., IEMOCAP, SEMAINE) is a lab-based video recording of subjects, with induced rather than posed affective states. More recently, physiological sensors have also been included, with various peripheral signals and EEG (e.g., DEAP, RECOLA, and MAHNOB-HCI [184]). A further approach is to use excerpts from movies and television shows, which can be labeled semi-automatically based on subtitles (e.g., AFEW and CASIA).

Another important aspect of databases is the employed model of affect. Of the 77 public databases used in the studies covered in our review, 58 use categorical models (75%), 14 use dimensional models (18%), and only 4 use both (5.2%). Further, only one database provides unlabeled affective displays (AUTOENCODER). This heavy reliance of databases on categorical models is also reflected in the models employed in the reviewed studies. Overall, 190 studies use categorical models (82%), 37 studies use dimensional models (16%), and only 6 use both (2.6%). For categorical affective states, every sequence is typically labeled in a discrete fashion with one affective state from a set of pre-defined labels; whereas, for dimensional affective states, frames are labeled continuously or in discrete steps. The ambiguity of human affect inherently makes both affect recognition and the labeling process difficult—there is an accuracy limit in the degree of agreement between multiple labelers. Overall, the trend apparent here goes towards capturing more naturalistic affective displays, as we venture from posed to spontaneous displays. Also, because of the challenges associated with categorical models, researchers have advocated for further investigating the application of dimensional models in affect recognition and comparing them with categorical models [9], [53], [56]. Unfortunately, at this stage, the number of examples per dataset does not see a clear upward trend yet and only few deep learning studies covered in our review investigate both types of models.

Unlabeled databases are used exclusively for transfer learning. When considering large general-purpose databases like ImageNet, the idea is to learn general low-level descriptors that help to improve initial model parameters. For FER, more relevant databases of unlabeled face images (e.g., VGG-Face) can be used. Smaller, labeled databases such as FER2013 are also used frequently for supervised pre-training. Note that these data sources for

TABLE 5
The top 15 most used databases in the reviewed studies; also included are three large databases published in 2016 and 2017.

Name	Year	Modality[a]			Examples[b]			Details on elicitation and annotation			Uses[c]	
		V	A	P	M	Subjects	Examples	Source	Annotation	Affect (Label)	Target	Transf.
CK+ [112]	2010	●			⊟	123	593	Posed (Lab)	Manual	Categorical (Discrete)	44	6
FER2013 [103]	2013	●			□		35887	Web search	Semi-aut.	Categorical (Discrete)	17	23
ImageNet [110]	2009	●			□		14.2M	Web search		(Generic image categories)	0	27
JAFFE [176]	1998	●			□	10	219	Posed (Lab)	Manual	Categorical (Discrete)	22	0
Emo-DB [164]	2005		●		⊟	10	800	Posed (Lab)	Manual	Categorical (Discrete)	17	3
VGG-Face [111]	2015	●			□	2622	2.6M	Web search		(Generic face images)	0	18
IEMOCAP [165]	2008	●	●		⊟	10	1039	Induced (Lab)	Manual	Cat./Dim. (Discrete)	14	3
SFEW2 [102]	2015	●			□		1635	Movies	Semi-aut.	Categorical (Discrete)	9	3
DEAP [177]	2012	●		●	⊟	32	40	Induced (Lab)	Semi-aut.	Dimensional (Discrete)	11	0
AFEW5 [178]	2015	●	●		⊟		1645	Movies	Semi-aut.	Categorical (Discrete)	9	0
eNTERFACE [179]	2005	●	●		⊟	42	1166	Induced (Lab)	Manual	Categorical (Discrete)	8	1
AFEW6 [12]	2016	●	●		⊟		1749	Movies	Semi-aut.	Categorical (Discrete)	8	0
RECOLA [180]	2013	●	●	●	⊟	23	46	Spont. (Lab)	Manual	Dimensional (Cont.)	8	0
CASIA [181]	2014	●	●		⊟	219	2 hr	TV shows	Manual	Categorical (Discrete)	7	1
SEMAINE [182]	2010	●	●		⊟	20	150	Induced (Lab)	Manual	Dimensional (Cont.)	7	0
AffectNet [57]	2017	●			□	450K	1M	Web search	Semi-aut.	Cat./Dim. (Discrete)	1	0
EmotioNet [183]	2016	●			□		1M	Web search	Automatic	Categorical (Discrete)	1	0
AUTOENCODER [62]	2017	●	●		⊟		6.5M	Web search		(Non-labeled affective displays)	0	1

[a] Modalities: V = Visual (FER, Body), A = Auditory (SER), P = Physiological (EEG, Peripheral, Other).
[b] M = Mode; □ = Static, ⊟ = Sequence; Number of subjects given where known.
[c] *Target* counts the number of studies that predicted given data; *Transfer* counts the number of studies that used given data for transfer learning.

transfer learning are primarily focused on static examples of the visual modality.

Databases published in 2016 and 2017 aim to provide sufficient training data for deep learning models. Compared to older databases, they contain many more examples, which is made possible by (semi-)automating the labeling process, or providing unlabeled examples. Three such databases are included in Table 5. Both AffectNet and EmotioNet are large web-based databases, each at around 1M labeled images. Notably, AffectNet includes both dimensional and categorical labels, encouraging studies to bridge the gap between affect representations. The AUTOENCODER dataset is the largest face video dataset with 6.5M examples; the dataset contains only 2777 labeled examples and is thus largely unlabeled. It can serve the purpose of unsupervised pre-training or semi-supervised learning [62].

In this review, we have thus far avoided directly comparing studies based on their recognition accuracies. Our reasoning is that even if both studies focus on the same dataset, differing test and training sets as well as evaluation statistics lead to difficulties in fairly comparing results. This situation is different for organized competitions, where the criteria are clearly defined and results independently verified. To illustrate state-of-the-art recognition accuracies, Table 6 reports the winning entries at affect recognition competitions from 2013 to 2017. These can generally be thought of as the latest and best results for the respective datasets at publication time. We employ the classification scheme introduced in Table 1 and used throughout this review to indicate the use of deep learning in these studies.

It is evident that since 2015, all winning entries at these competitions have made use of deep learning. Evaluation statistics are given by recognition accuracy for discrete affect recognition (EmotiW), and correlation coefficients for continuous affect recognition (AVEC). Progress in recognition

accuracies is measurable: On the AFEW dataset, which is extended with additional data every year, recognition accuracies have increased from 41% in 2013 to 60.3% in 2017.

4 DISCUSSION

In this review, we have seen that DNNs are part of most state-of-the-art affect recognition systems. They are applied for learning of (i) spatial feature representations, (ii) temporal feature representations, and (iii) joint feature representations for multimodal data. Considering the recent trend towards continuous and multimodal prediction of spontaneous affective displays in the wild, deep learning is generally well suited to address the challenges faced by such systems. Particularly, our results show that out of the 150 studies reporting comparisons between shallow and deep architectures, 95% reported that deep learning can lead to improvements over conventional approaches.

However, in comparison to related fields like object detection and ASR, the impact of deep learning has not yet been fully felt. This can in part be attributed to the higher difficulty and inherent ambiguity of affective displays—although many databases now provide annotations from multiple labelers, the ground truth tends to be unreliable and datasets are often imbalanced. A major obstacle is the relatively small size of labeled datasets. It hinders the ability of deep models to generalize well, and makes it difficult to train large models. In this vein, regularization techniques from general deep learning research [192] are one future research avenue. However, as of this writing, employed techniques such as knowledge transfer from related disciplines and data augmentation ultimately cannot fully make up for the lack of data.

Although we notice a trend towards larger labeled datasets, new approaches are needed to deal with this issue.

TABLE 6
Winning entries at affect recognition competitions (2013-2017).

Competition		Database	Winner	Deep feature learning			Evaluation statistic[b]
Name	Sub-Challenge[a]			Spatial	Temp.	Joint	
AVEC 2013	Fully cont. A/V	AVID	Meng et al. [185]				CC: 0.141
EmotiW 2013	A/V	AFEW3	Kahou et al. [186]	•			Accuracy: 41.0%
ICML 2013	Static FER	FER2013	Tang et al. [82]	•			Accuracy: 71.2%
INTERSPEECH'13	SER	GEMEP	Gosztolya et al. [187]				Accuracy: 73.5% (A), 63.3% (V)
AVEC 2014	Fully cont. A/V	AVID	Kachele et al. [188]				CC: 0.63 (A), 0.58 (V), 0.57 (D)
EmotiW 2014	A/V	AFEW4	Liu et al. [189]	•			Accuracy: 50.4%
AVEC 2015	Fully cont. A/V/P	RECOLA	He et al. [19]			•	CCC: 0.747 (A), 0.609 (V)
EmotiW 2015	A/V	AFEW5	Yao et al. [190]	•			Accuracy: 53.8%
EmotiW 2015	Static FER	SFEW2	Kim et al. [16]	•			Accuracy: 61.6%
AVEC 2016	Fully cont. A/V/P	RECOLA	Brady et al. [18]	•	•		CCC: 0.77 (A), 0.687 (V)
EmotiW 2016	A/V	AFEW6	Fan et al. [14]	•	•		Accuracy: 59.0%
EmotiW 2016	Group-level FER	HAPPEI	Li et al. [15]	•			RMSE: 0.82
AVEC 2017	Fully cont. A/V	SEWA	Chen et al. [17]	•	•	•	CCC: 0.68 (A), 0.76 (V), 0.51 (L)
EmotiW 2017	A/V	AFEW7	Hu et al. [95]	•			Accuracy: 60.3%
EmotiW 2017	Group-level FER	GAF	Tan et al. [191]	•			Accuracy: 80.9%

[a] Modalities: V = Visual (FER, Body), A = Auditory (SER), P = Physiological (EEG, Peripheral, Other).
[b] Reporting statistics used in competitions: CC = Correlation coefficient, CCC = Concordance correlation coefficient, A = Arousal, V = Valence, D = Dominance, L = Likability, RMSE = Root mean square error.

The labeling process for video-based affective datasets, especially with continuous annotations, is expensive and cannot easily be automated. Unsupervised and semi-supervised learning are promising trends in this regard. While unsupervised learning allows models to learn better initial parameters from unlabeled datasets, fine-tuning with labeled data is still required to "guide" the model towards its learning goal. The idea of semi-supervised learning is to label only a fraction of the examples, avoiding the expensive labeling process for the most part. These labeled examples can then be leveraged to learn from the remaining, much larger unlabeled part of the database that was acquired at a lower cost. Gupta et al. [62] demonstrated the feasibility of semi-supervised learning for affective computing, achieving promising results using 2777 labeled face videos to learn from the AUTOENCODER dataset of 6.5M unlabeled face video clips. In 2018, supervised pre-training was successfully used at Facebook [193] to transfer knowledge of 3.5B public Instagram images and hashtags for generic image classification and object detection. This development suggests that affect recognition too may benefit from much larger datasets. Unsupervised learning has also proved successful in SER–for example, Deng et al. [140] were able to improve the training process on labeled data by exploiting knowledge from unlabeled data using AEs.

In most state-of-the-art affect recognition models, handcrafted features still play an important role. While they are typically outperformed by DNN-learned features in direct comparison, some challenge-winning models rely on hybrid architectures that can take advantage of complementarities between the two [152], [18]. This can be attributed to multiple factors: Handcrafted features are readily available, widely recognized, and well-designed for specific applications such as FER and SER. On the other hand, deep models have not yet been established as methods of affect recognition. There are few specialized architectures researchers can draw from, which is why many fall back to ones from object recognition. Furthermore, the number of examples to draw from is not large enough to learn truly expressive deep features. Research suggests that the size of a dataset can be a bottleneck for performance of deep learning models [34]. As these circumstances change, larger and more expressive deep models specialized for affect recognition will have a chance of being established. Pre-trained deep models for FER and SER could become readily available as feature extractors, similar to handcrafted features.

As we have seen, affect recognition models are comprised of multiple components, including the steps of spatial and temporal feature learning as well as an additional fusion mechanism in the case of multimodal systems. We find that as of now, DNNs are generally applied in an isolated way to manage individual components—hence, a combination of multiple DNN-based components comes to mind. In fact, one ideal of classic deep learning research is the idea of integrating multiple components into globally trainable systems [194]. Some recent contributions have shown that affect recognition systems can combine multiple DNN-based components. As discussed throughout this review, the combination of CNN for feature learning and LSTM to learn the temporal context is most widely used (21 studies). However, most of these approaches train individual components separately (e.g., [106], [90]), which may lead to suboptimal performance considering the whole system. Future multimodal affect recognition systems are likely to be trained in an end-to-end fashion, which aligns training of the entire network closer with the true performance measure [195] but requires a larger amount of training examples. The feasibility and effectiveness of global end-to-end training has been demonstrated in two recent studies: Trigeorgis et al. [54] combined CNN and LSTM to learn features and temporal context directly from speech signals. Similarly, Tzirakis et al. [108] used the CNN-LSTM architecture to learn features and temporal context directly from raw audiovisual signals.

Another recent trend from deep learning research that lends itself to affect recognition of sequential data is the use of attention mechanisms [196]. They allow models to learn to pay more attention to promising segments of, say, a video
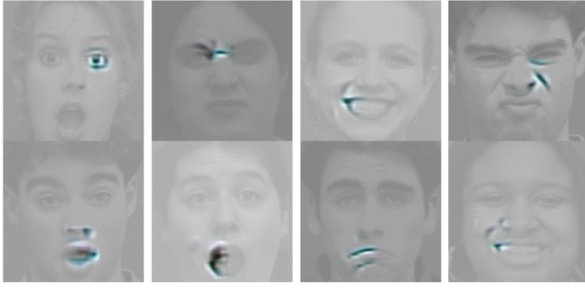
Fig. 7. CNNs learn high-level features similar to Action Units in FER (taken from [91] with permission).

of affective display. Mirsamadi et al. [63] found that even a simple attention mechanism contributes to SER by, for example, learning to ignore silent frames—an improvement that is easily interpretable and adds 1-2% to model accuracy.

In general, DNNs are often described as "black boxes", as their inner workings are complex and difficult to interpret. Recent research has attempted to improve this situation, for example by offering ways to visualize the activations of individual filters in CNNs [197]. Multiple studies have followed this approach to understand how CNNs recognize facial expressions, finding that higher layers learn concepts similar to Action Units (e.g., [87], [91]), validating previous research done in FER—Fig. 7 features some examples. Similarly, for SER, Tzirakis et al. [108] found that LSTM cells learn representations similar to well-known prosodic features. Such results indicate that the study of DNNs in affective computing could contribute to interdisciplinary emotion research by giving a further perspective on affective displays and their representation in general.

The question of categorical versus dimensional representations of affective states remains unsolved, as no deciding trend emerges from the literature. In practice, specific details of available modalities (e.g., static images or video data) and requirements regarding predictions (e.g., continuous or video-level) are relevant factors, and available labels on common databases widely dictate which representation is used. As can be seen in our review, the majority of studies still rely on categorical models of affect (82%), particularly in spatial feature learning (92%). However, given the nature of sensor data in typical HCI scenarios, it can be expected that dimensional representations, and combinations of categorical and dimensional models of affect will become even more relevant as a more natural way of dealing with continuous data [53]. This already becomes apparent in joint feature learning and to some extent in temporal feature learning, where dimensional models are already employed in 52% and 30% of the reviewed studies, respectively. Databases such as AffectNet [57], which provide categorical as well as dimensional annotations of affect in the wild, are an important step for furthering the investigation of dimensional models and for better integrating the context of affective display into human affect recognition.

ACKNOWLEDGMENTS

REFERENCES

[1] R. W. Picard, *Affective Computing*. MIT Press, 1997.
[2] ——, "Affective computing," MIT Media Lab., Tech. Rep., 1995.
[3] S. D'Mello, R. W. Picard, and A. Graesser, "Toward an affect-sensitive autotutor," *IEEE Intell. Syst.*, vol. 22, no. 4, pp. 53–61, 2007.
[4] C. Lisetti, F. Nasoz, C. LeRouge, O. Ozyer, and K. Alvarez, "Developing multimodal intelligent affective interfaces for tele-home health care," *Int. J. Hum.-Comput. Stud.*, vol. 59, no. 1, pp. 245–255, 2003.
[5] G. N. Yannakakis and J. Togelius, "Experience-driven procedural content generation," *IEEE Trans. Affect. Comput.*, vol. 2, no. 3, pp. 147–161, 2011.
[6] F. De Rosis, C. Pelachaud, I. Poggi, V. Carofiglio, and B. De Carolis, "From greta's mind to her face: Modelling the dynamics of affective states in a conversational embodied agent," *Int. J. Hum.-Comput. Stud.*, vol. 59, no. 1, pp. 81–118, 2003.
[7] M. Pantic and L. J. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proc. IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.
[8] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, 2001.
[9] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, 2010.
[10] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Comput. Vision and Image Understanding*, vol. 108, no. 1, pp. 116–134, 2007.
[11] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.
[12] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon, "Emotiw 2016: Video and group-level emotion recognition challenges," in *Proc. Int. Conf. Multimodal Interact.*, 2016, pp. 427–432.
[13] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. AVEC*, 2016, pp. 3–10.
[14] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proc. Int. Conf. Multimodal Interact.*, 2016, pp. 445–450.
[15] J. Li, S. Roy, J. Feng, and T. Sim, "Happiness level prediction with sequential inputs via multiple regressions," in *Proc. Int. Conf. Multimodal Interact.*, 2016, pp. 487–493.
[16] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee, "Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition," in *Proc. Int. Conf. Multimodal Interact.*, 2015, pp. 427–434.
[17] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proc. AVEC*, 2017, pp. 19–26.
[18] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *Proc. AVEC*, 2016, pp. 97–104.
[19] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proc. AVEC*, 2015, pp. 73–80.
[20] Y. Bengio and Y. LeCun, "Scaling learning algorithms towards ai," in *Large-Scale Kernel Machines*, L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, Eds. MIT Press, 2007, pp. 1–41.
[21] Y. Bengio, O. Delalleau, and N. Le Roux, "The curse of dimensionality for local kernel machines," Université de Montréal, Tech. Rep., March 2005.
[22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
[23] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
[25] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[26] Y. Bengio, "Learning deep architectures for ai," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[27] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[30] J. Martens and I. Sutskever, "Learning recurrent neural networks with hessian-free optimization," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 1033–1040.

[31] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2011, pp. 315–323.

[32] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1139–1147.

[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Mach. Learn. Res.*, vol. 37, 2015, pp. 448–456.

[34] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 843–852.

[35] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, no. 1, pp. 106–154, 1962.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, vol. 25, 2012, pp. 1097–1105.

[37] Y. LeCun, L. Jackel, B. Boser, J. Denker, H. Graf, I. Guyon, D. Henderson, R. Howard, and W. Hubbard, "Handwritten digit recognition: Applications of neural network chips and automatic learning," *IEEE Commun. Mag.*, vol. 27, no. 11, pp. 41–46, 1989.

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F.-F. Li, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[39] Y. Ming, S. Cao, R. Zhang, Z. Li, Y. Song, and H. Qu, "Understanding hidden memories of recurrent neural networks," in *Proc. Conf. Visual Anal. Sci. Technol.*, 2017.

[40] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *arXiv preprint arXiv:1506.00019*, 2015.

[41] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2017, pp. 1724–1734.

[42] K. J. Han, A. Chandrashekaran, J. Kim, and I. Lane, "Densely connected networks for conversational speech recognition," in *Proc. INTERSPEECH*, 2018, pp. 796–800.

[43] P. Baldi and K. Hornik, "Neural networks and principal component analysis: Learning from examples without local minima," *Neural Networks*, vol. 2, pp. 53–58, 1989.

[44] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," in *Parallel Distributed Processing: Volume 1: Foundations*, D. E. Rumelhart and J. L. McClelland, Eds. MIT Press, 1986, pp. 194–281.

[45] R. Salakhutdinov and H. Larochelle, "Efficient learning of deep boltzmann machines," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2010, pp. 693–700.

[46] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. NIPS*, vol. 19, 2007, pp. 153–160.

[47] M. Pantic, N. Sebe, J. F. Cohn, and T. Huang, "Affective multimodal human-computer interaction," in *Proc. Int. Conf. Multimedia*, 2005, pp. 669–676.

[48] P. Kragel and K. LaBar, "Decoding the nature of emotion in the brain," *Trends in Cogn. Sciences*, vol. 20, no. 6, pp. 444–455, 2016.

[49] A. Ortony and T. J. Turner, "What's basic about basic emotions?" *Psychological Review*, vol. 97, no. 3, pp. 315–331, 1990.

[50] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Semiotica*, vol. 1, no. 1, pp. 49–98, 1969.

[51] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of Emotion*, 1980, pp. 3–33.

[52] W. James, *The Principles of Psychology, Volume I.* Holt, 1890.

[53] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image Vision Comput.*, vol. 31, pp. 120–136, 2013.

[54] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. ICASSP*, 2016, pp. 5200–5204.

[55] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, and B. Hu, "Emotion recognition from multi-channel eeg data through convolutional recurrent neural network," in *Proc. Int. Conf. Bioinf. and Biomed.*, 2016, pp. 352–359.

[56] L. A. Bugnon, R. A. Calvo, and D. H. Milone, "Dimensional affect recognition from hrv: An approach based on supervised som and elm," *IEEE Trans. Affect. Comput.*, 2017.

[57] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, 2017.

[58] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[59] S. K. D'Mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surveys*, vol. 47, no. 3, p. 43, 2015.

[60] S. J. Pan, Q. Yang *et al.*, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[61] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *Proc. Europ. Conf. Comput. Vision*, 2012, pp. 808–822.

[62] O. Gupta, D. Raviv, and R. Raskar, "Multi-velocity neural networks for facial expression recognition in videos," *IEEE Trans. Affect. Comput.*, 2017.

[63] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. ICASSP*, 2017, pp. 2227–2231.

[64] D. H. Kim, W. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Trans. Affect. Comput.*, 2017.

[65] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognit. Lett.*, vol. 66, pp. 22–30, 2015.

[66] M. Chen, L. Zhang, and J. P. Allebach, "Learning deep features for image emotion classification," in *Proc. Int. Conf. Image Process.*, 2015, pp. 4491–4495.

[67] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proc. Int. Soc. Music Inf. Retrieval*, 2010, pp. 255–266.

[68] E. M. Schmidt and Y. E. Kim, "Learning emotion-based acoustic features with deep belief networks," in *Proc. Worksh. Appl. Signal Process. Audio Acoust.*, 2011, pp. 65–68.

[69] R. A. Calvo and S. Mac Kim, "Emotions in text: Dimensional and categorical models," *Comput. Intell.*, vol. 29, no. 3, pp. 527–543, 2013.

[70] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, 2001.

[71] A. Mehrabian, "Communication without words," *Psychology Today*, vol. 2, no. 4, pp. 53–56, 1968.

[72] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, 2015.

[73] C. A. Corneanu, M. O. Simon, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1548–1568, 2016.

[74] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proc. Int. Conf. Multimodal Interact.*, 2015, pp. 503–510.

[75] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, "Holonet: Towards robust emotion recognition in the wild," in *Proc. Int. Conf. Multimodal Interact.*, 2016, pp. 472–478.

[76] Y. Cheng, B. Jiang, and K. Jia, "A deep structure for facial expression recognition under partial occlusion," in *Proc. Int. Conf. Intell. Inf. Hiding Multim. Sign. Proc.*, 2014, pp. 211–214.

[77] Y. Lv, Z. Feng, and C. Xu, "Facial expression recognition via deep learning," in *Proc. SMARTCOMP*, 2014, pp. 303–308.

[78] C. Fadil, R. Alvarez, C. Martínez, J. Goddard, and H. Rufiner, "Multimodal emotion recognition using deep networks," in *Latin Am. Congr. Biomed. Eng.*, 2014, pp. 813–816.

[79] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[80] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2014.

[81] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. CVPR*, 2015, pp. 1–9.

[82] Y. Tang, "Deep learning using linear support vector machines," in *Proc. Int. Conf. Mach. Learn.*, 2013.

[83] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, 2013.

[84] P. Barros, D. Jirak, C. Weber, and S. Wermter, "Multimodal emotional state recognition using sequence-dependent deep hierarchical features," *Neural Networks*, vol. 72, pp. 140–151, 2015.

[85] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio-visual emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, 2017.

[86] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *Proc. Int. Conf. Multimodal Interact.*, 2017, pp. 524–528.

[87] P. Khorrami, T. Paine, and T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *Proc. Int. Conf. Comput. Vision*, 2015, pp. 19–27.

[88] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. Int. Conf. Comput. Vision*, 2015, pp. 2983–2991.

[89] S. Chen, X. Li, Q. Jin, S. Zhang, and Y. Qin, "Video emotion recognition in the wild based on fusion of multimodal features," in *Proc. Int. Conf. Multimodal Interact.*, 2016, pp. 494–500.

[90] P. Khorrami, T. Le Paine, K. Brady, C. Dagli, and T. Huang, "How deep neural networks can improve emotion recognition on video data," in *Proc. Int. Conf. Image Process.*, 2016, pp. 619–623.

[91] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," *arXiv preprint arXiv:1705.01842*, 2017.

[92] H. Li, J. Sun, Z. Xu, and L. Chen, "Multimodal 2d+3d facial expression recognition with deep fusion convolutional neural network," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2816–2831, 2017.

[93] L. Gui, T. Baltrušaitis, and L.-P. Morency, "Curriculum learning for facial expression recognition," in *Proc. Int. Conf. Automat. Face Gesture Recognit.*, 2017, pp. 505–511.

[94] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, 2017.

[95] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proc. Int. Conf. Multimodal Interact.*, 2017, pp. 553–560.

[96] M. M. Ghazi and H. K. Ekenel, "Automatic emotion recognition in the wild using an ensemble of static and dynamic representations," in *Proc. Int. Conf. Multimodal Interact.*, 2016, pp. 514–521.

[97] W. Ding, M. Xu, D. Huang, W. Lin, M. Dong, X. Yu, and H. Li, "Audio and face video emotion recognition in the wild using deep neural networks and small datasets," in *Proc. Int. Conf. Multimodal Interact.*, 2016, pp. 506–513.

[98] B. Sun, L. Li, G. Zhou, and J. He, "Facial expression recognition in the wild based on multimodal texture features," *J. Electron. Imaging*, vol. 25, no. 6, pp. 1–8, 2016.

[99] H. Kaya, F. Gürpınar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image Vision Comput.*, vol. 65, pp. 66–75, 2017.

[100] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Video emotion recognition with transferred deep feature encodings," in *Proc. Int. Conf. Multimedia Retrieval*, 2016, pp. 15–22.

[101] J. Susskind, A. Anderson, and G. E. Hinton, "The toronto face dataset," University of Toronto, Tech. Rep. TR 2010-001, 2010.

[102] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proc. Int. Conf. Comput. Vision Workshops*, 2011, pp. 2106–2112.

[103] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, 2015.

[104] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proc. CVPR*, 2015, pp. 4295–4304.

[105] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE Trans. Cybern.*, vol. PP, no. 99, pp. 1–11, 2017.

[106] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proc. Int. Conf. Multimodal Interact.*, 2015, pp. 467–474.

[107] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proc. Int. Conf. Multimodal Interact.*, 2015, pp. 443–449.

[108] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 8, pp. 1301–1309, 2017.

[109] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *Proc. CVPR Workshops*, 2014, pp. 806–813.

[110] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.

[111] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition," in *Proc. Brit. Mach. Vision Conf.*, 2015, pp. 1–12.

[112] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Proc. CVPR Workshops*, 2010, pp. 94–101.

[113] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, and D. Tao, "Deep neural networks with relativity learning for facial expression recognition," in *Proc. Int. Conf. Multimedia Expo Workshops*, 2016, pp. 1–6.

[114] Y. Cai, W. Zheng, T. Zhang, Q. Li, Z. Cui, and J. Ye, "Video based emotion recognition using cnn and brnn," in *Proc. Chin. Conf. Pattern Recognition*, 2016, pp. 679–691.

[115] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in *Proc. Int. Conf. Automat. Face Gesture Recognit.*, 2017, pp. 118–126.

[116] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[117] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[118] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," *Artif. Intell. Rev.*, vol. 43, no. 2, pp. 155–177, 2015.

[119] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. NIPS*, 2009, pp. 1096–1104.

[120] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, 2014.

[121] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016.

[122] F. Eyben, M. Wöllmer, and B. Schuller, "Openear—introducing the munich open-source emotion and affect recognition toolkit," in *Proc. ACII*, 2009, pp. 1–6.

[123] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge," in *Proc. INTERSPEECH*, 2013.

[124] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[125] L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, and H. Sahli, "Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition," in *Proc. ACII*, 2013, pp. 312–317.

[126] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proc. Int. Conf. Multimedia*, 2014, pp. 801–804.

[127] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proc. Int. Conf. Platform Technol. Serv.*, 2017, pp. 1–5.

[128] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[129] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," *Image Vision Comput.*, vol. 31, no. 2, pp. 153–163, 2013.

[130] H. M. Fayek, M. Lech, and L. Cavedon, "Towards real-time speech emotion recognition using deep neural networks," in *Proc. Int. Conf. Signal Process. Commun. Syst.*, 2015, pp. 1–5.

[131] W. Zheng, J. Yu, and Y. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," in *Proc. ACII*, 2015, pp. 827–831.

[132] C.-T. Ho, Y.-H. Lin, and J.-L. Wu, "Emotion prediction from user-generated videos by emotion wheel guided deep learning," in *Proc. Int. Conf. Neural Inf. Process.*, 2016, pp. 3–12.

[133] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines," in *Proc. ICASSP*, 2011, pp. 5884–5887.

[134] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *Proc. ICASSP*, 2017, pp. 5115–5119.

[135] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *Proc. ICASSP*, 2017, pp. 4990–4994.

[136] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. Provost, "Progressive neural networks for transfer learning in emotion recognition," in *Proc. INTERSPEECH*, 2017, pp. 1098–1102.

[137] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proc. ACII*, 2013, pp. 511–516.

[138] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Proc. NIPS*, 2016.

[139] S. Pini, O. B. Ahmed, M. Cornia, L. Baraldi, R. Cucchiara, and B. Huet, "Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild," in *Proc. Int. Conf. Multimodal Interact.* ACM, 2017, pp. 536–543.

[140] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 500–504, 2017.

[141] E. Coutinho, J. Deng, and B. Schuller, "Transfer learning emotion manifestation across music and speech," in *Proc. Int. Joint Conf. Neural Networks*, 2014, pp. 3592–3598.

[142] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *Proc. ICASSP*, 2017.

[143] S. H. Fairclough, "Fundamentals of physiological computing," *Interact. Comput.*, vol. 21, no. 1-2, pp. 133–145, 2008.

[144] F. H. Wilhelm and P. Grossman, "Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment," *Biol. Psychol.*, vol. 84, no. 3, pp. 552–569, 2010.

[145] M. Yanagimoto and C. Sugimoto, "Recognition of persisting emotional valence from eeg using convolutional neural networks," in *Proc. Int. Workshop Computat. Intell. Appl.*, 2016, pp. 27–32.

[146] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial-temporal recurrent neural network for emotion recognition," *IEEE Trans. Cybern.*, 2017.

[147] J. Li, Z. Zhang, and H. He, "Implementation of eeg emotion recognition system based on hierarchical convolutional neural networks," in *Proc. Int. Conf. Brain Inspired Cogn. Syst.*, 2016, pp. 22–33.

[148] B. Zhang, G. Essl, and E. Mower Provost, "Automatic recognition of self-reported and perceived emotion: Does joint modeling help?" in *Proc. Int. Conf. Multimodal Interact.*, 2016, pp. 217–224.

[149] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *Proc. Winter Conf. Appl. Comput. Vision*, 2016, pp. 1–9.

[150] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affect. Comput.*, vol. 2, no. 2, pp. 92–105, 2011.

[151] J. Wei, E. Pei, D. Jiang, H. Sahli, L. Xie, and Z. Fu, "Multimodal continuous affect recognition based on lstm and multiple kernel learning," in *Proc. Annu. Summit Conf. Asia-Pac. Signal Inf. Proc. Assoc.*, 2014, pp. 1–4.

[152] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, Y. Zong, and N. Sun, "Multi-clue fusion for emotion recognition in the wild," in *Proc. Int. Conf. Multimodal Interact.*, 2016, pp. 458–463.

[153] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. CVPR*, 2014, pp. 1725–1732.

[154] K. Kaza, A. Psaltis, K. Stefanidis, K. C. Apostolakis, S. Thermos, K. Dimitropoulos, and P. Daras, "Body motion analysis for emotion recognition in serious games," in *Proc. Int. Conf. Universal Access in Hum.-Comput. Interact.*, 2016, pp. 33–42.

[155] M. Gales, S. Young *et al.*, "The application of hidden markov models in speech recognition," *Found. Trends Signal Process.*, vol. 1, no. 3, pp. 195–304, 2008.

[156] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*, 2013, pp. 6645–6649.

[157] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. INTERSPEECH*, 2009.

[158] B. W. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "The interspeech 2012 speaker trait challenge," in *Proc. INTERSPEECH*, 2012, pp. 254–257.

[159] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. ICASSP*, 2013, pp. 3687–3691.

[160] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski *et al.*, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[161] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. ICASSP*, 2011, pp. 5688–5691.

[162] F. Lingenfelser, J. Wagner, J. Deng, R. Bruckner, B. Schuller, and E. Andre, "Asynchronous and event-based fusion systems for affect recognition on naturalistic data in comparison to conventional approaches," *IEEE Trans. Affect. Comput.*, 2016.

[163] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Proc. Annu. Summit Conf. Asia-Pac. Signal Inf. Proc. Assoc.*, 2016, pp. 1–4.

[164] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. INTERSPEECH*, vol. 5, 2005, pp. 1517–1520.

[165] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources Eval.*, vol. 42, no. 4, p. 335, 2008.

[166] S. Jirayucharoensak, S. Pan-Ngum, and P. Israsena, "Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation," *The Scientific World Journal*, vol. 2014, 2014.

[167] W.-L. Zheng, J.-Y. Zhu, Y. Peng, and B.-L. Lu, "Eeg-based emotion classification using deep belief networks," in *Proc. Int. Conf. Multimedia Expo*, 2014, pp. 1–6.

[168] H. Xu and K. N. Plataniotis, "Affective states classification using eeg and semi-supervised deep learning approaches," in *Proc. Int. Workshop Multimedia Signal Process.*, 2016, pp. 1–6.

[169] R. M. Mehmood, R. Du, and H. J. Lee, "Optimal feature selection and deep learning ensembles method for emotion recognition from human brain eeg sensors," *IEEE Access*, vol. 5, pp. 14 797–14 806, 2017.

[170] L. Bozhkov, P. Koprinkova-Hristova, and P. Georgieva, "Learning to decode human emotions with echo state networks," *Neural Networks*, vol. 78, pp. 112–119, 2016.

[171] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang, "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model," *Comput. Meth. Programs Biomed.*, vol. 140, pp. 93–110, 2017.

[172] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *IEEE Comput. Intell. Mag.*, vol. 8, no. 2, pp. 20–33, 2013.

[173] F. Lingenfelser, J. Wagner, and E. André, "A systematic discussion of fusion techniques for multi-modal affect recognition tasks," in *Proc. Int. Conf. Multimodal Interact.*, 2011, pp. 19–26.

[174] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, 2017.

[175] W. Liu, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal deep learning," in *Proc. Int. Conf. Neural Inf. Process.*, 2016, pp. 521–529.

[176] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. Int. Conf. Automat. Face Gesture Recognit.*, 1998, pp. 200–205.

[177] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, 2012.

[178] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *Proc. Int. Conf. Multimodal Interact.*, 2015, pp. 423–426.

[179] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. Int. Conf. Data Eng. Workshops*, 2006, pp. 8–8.

[180] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Proc. Int. Conf. Automat. Face Gesture Recognit.*, 2013, pp. 1–8.

[181] W. Bao, Y. Li, M. Gu, M. Yang, H. Li, L. Chao, and J. Tao, "Building a chinese natural emotional audio-visual database," in *Proc. Int. Conf. Signal Process.*, 2014, pp. 583–587.

[182] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 5–17, 2012.

[183] C. F. Benitez-Quiroz, R. Srinivasan, A. M. Martinez *et al.*, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proc. CVPR*, 2016, pp. 5562–5570.

[184] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, 2012.

[185] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proc. AVEC*, 2013, pp. 21–30.

[186] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proc. Int. Conf. Multimodal Interact.*, 2013, pp. 543–550.

[187] G. Gosztolya, R. Busa-Fekete, and L. Tóth, "Detecting autism, emotions and social signals using adaboost," in *Proc. INTERSPEECH*, 2013, pp. 220–224.

[188] M. Kächele, M. Schels, and F. Schwenker, "Inferring depression and affect from application dependent meta knowledge," in *Proc. AVEC*, 2014, pp. 41–48.

[189] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *Proc. Int. Conf. Multimodal Interact.*, 2014, pp. 494–501.

[190] A. Yao, J. Shao, N. Ma, and Y. Chen, "Capturing au-aware facial features and their latent relations for emotion recognition in the wild," in *Proc. Int. Conf. Multimodal Interact.*, 2015, pp. 451–458.

[191] L. Tan, K. Zhang, K. Wang, X. Zeng, X. Peng, and Y. Qiao, "Group emotion recognition with individual facial emotion cnns and global image based cnns," in *Proc. Int. Conf. Multimodal Interact.*, 2017, pp. 549–552.

[192] J. Kukačka, V. Golkov, and D. Cremers, "Regularization for deep learning: A taxonomy," *arXiv preprint arXiv:1710.10686*, 2017.

[193] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proc. Europ. Conf. Comput. Vision*, 2018, pp. 185–201.

[194] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[195] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1764–1772.

[196] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017.

[197] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Europ. Conf. Comput. Vision*, 2014, pp. 818–833.

Philipp V. Rouast received the B.Sc. and M.Sc. degrees in Industrial Engineering from Karlsruhe Institute of Technology, Germany, in 2013 and 2016 respectively. He is currently working towards the Ph.D. degree in Information Systems and is a graduate research assistant at The University of Newcastle, Australia. His research interests include deep learning, affective computing, HCI, and related applications of computer vision. Find him at https://prouast.github.io.

Marc T. P. Adam is a Senior Lecturer in Computing and Information Technology at the University of Newcastle, Australia. In his research, he investigates the interplay of human users' cognition and affect in human-computer interaction. He received an undergraduate degree in Computer Science from the University of Applied Sciences Würzburg, Germany, and a PhD in Economics of Information Systems from Karlsruhe Institute of Technology, Germany.

Raymond Chiong (M'05-SM'14) is a Senior Lecturer at the University of Newcastle, Australia. He received his MSc from the University of Birmingham, UK, and PhD from the University of Melbourne, Australia. His research interests include evolutionary game theory, optimization, data analytics, and modeling of complex adaptive systems. He is the Editor-in-Chief of the Journal of Systems and Information Technology, an Editor of Engineering Applications of Artificial Intelligence, and an Associate Editor of the Computational Intelligence Magazine.