Internet of emotional people: Towards continual affective computing cross cultures via audiovisual signals

Jing Han [a,*], Zixing Zhang [b], Maja Pantic [c], Björn Schuller [a,b]

[a] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany
[b] GLAM – the Group on Language, Audio & Music, Imperial College London, London SW7 2AZ, UK
[c] Intelligent Behaviour Understanding Group, Imperial College London, London SW7 2AZ, UK

1. Introduction

Recently, a new Internet paradigm called "Internet of People (IoP)" has emerged for promoting the current Internet of Thing frameworks into the next generation that is more personal, user-centric, human-driven [1,2]. In this context, plenty of research directions need to be investigated to close the gap between human intelligence and machine intelligence and therefore accomplish user-centric cyber-physical-social systems [3–5]. These research directions include but are not limited to crowdsensing [6], financial data protection [7], health monitoring [8], and behaviour analysis [9].

Moreover, sentiment analysis and affective computing plays an essential role in enabling machines emotional intelligence, which itself is an interdisciplinary research area spanning various research fields such as computer science, cognitive science, and social science [10–15]. With the usage popularity of intelligent edge devices, such as smartphones, wearable rings or watches, smart speakers, intelligent vehicles, automatically detecting and analysing human's emotional or affective states via the Internet becomes more attractive and feasible, which is defined as *Internet of Emotional People* (IoEP) henceforth. The importance of IoEP is at least twofold: (i) It enables the intelligent assistants in devices to be more friendly and empathetic, and consequently promotes their user experience. (ii) It particularly benefits for patients, who are either suffering from a mental health disorder or in rehabilitation progress by the disease, such as stress, anxiety, and depression. By tracking their mental states, it helps the doctor to make a diagnostic scheme and provide therapies for these mental and emotional problems accordingly.

Albeit the interest and attractiveness of IoEP, there are two essential characteristics, i. e., diversity and scarcity. The *diversity* suggests that the information of interest around humans is divergent for different groups and variable in different time periods. In this context, it is of necessity to devise individual models for the tasks of interest to achieve appropriate performance. Such a conventional isolated learning strategy, however, suffers from several fundamental issues, which severely prevents us from integrating affective computing technologies into real human-centric applications. It is time and expert prohibitive to manually and repeatedly design separate models for each task. Although this issue can be partially solved by certain technologies like automated Machine Learning (autoML) that lets machines automatically construct the model structure without exhausting

* Corresponding author.
E-mail address: jing.han@informatik.uni-augsburg.de (J. Han).

human interference [16,17], it is somewhat pragmatically unfeasible to implement so many models into one system due to the limited memory size, computational resources, and energy. Whereas, the *scarcity* characteristic of IoEP implies that only a limited amount of labelled data can be used in many individual training cases, due to the time- and cost-consuming annotation process [18]. Without sufficient training data, it is hard to obtain a reliable and robust model, especially for a neural network-based model owing to its data-hungry essence. These two characteristics severely hinder the deployment progress of IoEP, because conventional learning strategies generally tackle each task *independently*, and largely overlook the knowledge shared among different tasks.

To cope with these issues, the most popular strategies are associated with *Transfer Learning* (TL) [18] and *Multi-Task Learning* (MTL) [19–21]. With TL, a pre-trained model for a task with a rich resource can be reused for another recognition task with low resource [22]. Thus, TL can well address the data scarcity problem, and relax the task-dependent assumption to a certain extent. Nevertheless, it is still particularly designed for improving the target task, and suffers a serious *catastrophic forgetting* — a phenomenon that the model endures an abrupt performance decrease or, in the worst case, being completely overwritten by the new task [23,24]. In comparison with TL, MTL learns several tasks by one model simultaneously. Therefore, it cannot only largely reduce the number of models, but also efficiently exploit the shared information among various tasks. However, in the training stage, it requires a large size of training data, and consequently needs large storage memory size and heavy computational load. Even worse, when facing a new task, the model has to be re-trained from scratch, which significantly reduces its flexibility [23,24].

Motivated by these analyses, in the present article we intend to shed some fresh light on affective computing by using a *lifelong learning* paradigm, with the purpose of renovating the current *isolated* learning strategy into the next *continual* generation, namely *continual affective computing*. Lifelong learning, also known as *continual learning*, was initially proposed by Thrun and Mitchell in 1995 for a robot control [25]. It aims to empower machines with the capability of continually acquiring and transferring knowledge and skills throughout their lifespan, just like human beings. That is, concepts and their relationships learnt in the past can help us understand and learn a new subject better, because a lot of information and knowledge are shared across domains and tasks [23–25]. Therefore, with lifelong learning, a machine can retain and accumulate the knowledge learnt in the past and make use of the knowledge seamlessly in future learning [23,24].

Note that, lifelong learning differs from TL, where the learning process is one-directional — previous tasks help the current task, but not the other way around. Likewise, it is distinct from MTL that is absolutely not continual. Generally speaking, lifelong learning holds multifold advantages [23,24]: (i) Continuous learning. It avoids the requirement of a large number of models when dealing with many different tasks sequentially. (ii) Knowledge accumulation and transformation. It can accumulate knowledge from previous tasks, and then use it to help with new task learning. Hence, it efficiently makes use of the relationship among different tasks, and partially reduces the needs of a large amount of training data for each task. (iii) Discovery of new tasks. That is, the model with lifelong learning can learn in a dynamic open environment by itself where novel objects and scenarios that have not been learnt before may be available. Because of these advantages, lifelong learning has emerged as one of the leading potentials to handle the aforementioned problems raised for affective computing.

To the best of our knowledge, merely a handful of related studies are available in the affective computing literature [26, 27]. These studies, nonetheless, merely focused on the sentiment analysis in the context of natural language processing [26,27]. It remains unclear how it performs for many other tasks with other cues, such as acoustic or visual signals. For this reason, we, for the first time, apply lifelong learning to audiovisual affective computing, taking the example of emotion recognition in a cross-culture scenario, i. e., cross the tasks of *French emotion recognition* and *German emotion recognition* in our work. The rationale behind this task selection is threefold: (i) Audiovisual emotion recognition is one of the most active research domains in machine learning nowadays [28,29]. (ii) It has been systematically benchmarked by a series of challenges [30]. (iii) Additionally, it has been widely known that recognising emotions across cultures is still a major challenge, given the variety of languages and cultural backgrounds [31–33].

Furthermore, in this work, we introduce the lifelong learning algorithm of *Elastic Weight Consolidation* (EWC) for the raised task [34]. As the first work in this direction, we attempt to investigate the feasibility of the proposed lifelong learning for affective computing in the audio and/or video modalities, in an application of cross-cultural emotion recognition. Finally, we conduct extensive experiments on two emotional databases, which were recorded in different cultures, i. e., French and German. We further report results on the two cultures, and visualise the performance improvement and effect of EWC for the tasks at hand, followed by a detailed analysis.

The remainder of this article is structured as follows. In Section 2, we briefly review past and related studies. Then, in Section 3, we present a detailed description of the introduced lifelong learning and EWC, as well as its learning process in cross-cultural audiovisual emotion recognition. After that, we introduce the selected databases, features, and experimental setups in Section 4, before delivering detailed experimental results and discussions in Section 5. Finally, we draw our conclusions and highlight future research directions in Section 6.

2. Related work

As mentioned in Section 1, TL and MTL are widely considered to be two of the most frequently used strategies to address the cross-task problem in the context of affective computing. Plenty of related studies are available in the literature. Hereinafter, we concentrate the investigation on emotion or sentiment recognition.

2.1. Transfer learning

The principle of TF is to transfer knowledge from a source task or domain to a target one. The tasks or domains are generally determined by the discrepancies in the aspects of modalities, data distributions, and label spaces [18,22,35]. Specifically, to transfer the knowledge *across modalities*, Albantie et al. [36] proposed a teacher–student learning framework, where a speech emotion recognition network (the student) is trained by distilling the knowledge of a pre-trained facial emotion recognition network (the teacher) across unlabelled videos. This framework, however, assumes the availability of a well-trained complex teacher network. To relax this assumption, Han et al. [37] introduced an emotion embedding framework across multiple modalities. It considered a triplet loss to minimise the distance of the intra-class samples while maximising the distance of the inter-class samples, regardless of the modalities where the samples come from. In doing so, the salient emotional information is able to flow freely across modalities.

To reuse the knowledge from *mismatched data distributions*, Glorot et al. [38] firstly used a stacked denoising autoencoder for

sentiment classification. By building the model in a hierarchical and an unsupervised manner, it is able to discover intermediate text representations of the review comments for different products. After that, similar studies have been done for cross-corpus speech emotion recognition [39,40]. Rather than extracting the high-level representative features, domain-invariant acoustic representations by means of a domain adversarial training approach was further proposed and investigated as well recently [41–43]. By this method, an additional network is trained to distinguish the domains from which the extracted representations come.

Differ from the TL algorithms designed to mitigate the modality and feature mismatch, fine-tuning seems to be the most frequently used way to deal with the *label mismatched* problem. For example, the studies [44,45] have shown that the models pre-trained on large-scale datasets for other classification tasks can be fine-tuned on emotion recognition tasks to learn acoustic and visual representations. A similar investigation was shown for sentiment analysis as well [46].

Nonetheless, all aforementioned TL studies merely focus on improving the model performance on the following tasks and ignore how it performs on previous ones. Therefore, it fails to address the challenges when facing diverse and changeable tasks as demonstrated in Section 1.

2.2. Multi-Task Learning

MTL is frequently utilised to reduce the number of models while it largely explores the shared information across all tasks. For speech emotion recognition, Eyben et al. [19] firstly proposed to jointly train five different emotional dimensions. The experimental results have indicated that the MTL model remarkably outperforms single-task-based models. Following this work, Han et al. [47] combined the emotion prediction with an annotation uncertainty as joint tasks to be learnt together. Moreover, Xia and Liu [48] suggested incorporating the losses from both the categorical and the dimensional emotion recognition to optimise the neural networks. Further, Zhang et al. [21] investigated MTL in a cross-corpus scenario, where many tasks, such as corpus, domain, and gender distinctions, were considered to be optimised along with emotion recognition. Other similar studies have also been reported in [49–51].

When compared with TL, MTL is capable of learning multiple tasks simultaneously. Nevertheless, it requires to retrain the model from scratch when an unseen task comes. Hence, MTL is very computationally inefficient, and hinders the learning of novel tasks in real time [23,24]. This further highlights the necessity of lifelong learning for affective computing in IoEP applications.

In our study, to evaluate the performance of the proposed lifelong learning paradigm, we take MTL as a performance upper bound and refer it to be *joint training*. That is, training data intended for different tasks are agglomerated as one larger dataset in the training process.

2.3. Emotion recognition in general

In affective computing, emotion recognition can be defined as a process of automatically perceiving the affect of human beings, and it usually leverages methodologies and techniques from multiple research areas covering signal processing, machine learning, and so on. In particular, two kinds of emotion models are frequently explored, namely, categorical models and dimensional models. That is, the perceived emotions can be presented as discrete labels from multiple discrete categories such as happy and sad [52], or continuous values in dimensional spaces such as the dimensions of arousal (the degree of intensity of an emotional state) and valence (how positive or negative an emotional state is) in a circular representation [53].

With that said, different systems and frameworks have been proposed to recognise a person's emotional state from modalities such as audio, text, video, and physiological signals. Alternatively, one can also build multimodal systems to integrate information from different modalities [28,54,55]. Moreover, similarly to other applications where deep neural networks are employed, advances in emotion recognition have also benefited from deep learning techniques. For a comprehensive overview and in-depth discussion of the state-of-the-art, remaining challenges and open issues in emotion recognition, readers are referred to [10,13,56,57].

3. Lifelong learning for audiovisual emotion recognition

As a prospective study of continual affective computing, in this paper, we carry out a cross-cultural audiovisual emotion recognition task as a first attempt, to demonstrate how lifelong learning can be achieved in affective computing domain. For this reason, in the following sections, we first briefly describe the definition and goals of lifelong learning in Section 3.1. After that, a typical lifelong learning paradigm is presented in Section 3.2, which aims to lessen the catastrophic forgetting effect when training a model on multiple tasks sequentially. Then, procedures on how to construct and train a continual emotion recognition model are elaborated in detail in Section 3.3.

3.1. Lifelong learning and catastrophic forgetting

Lifelong learning was first defined in [25] as a learning algorithm applied in a lifelong context, where a series of tasks can be learnt sequentially instead of in isolation, so that knowledge can be transferred across these tasks. In particular, like humans, knowledge obtained in previous learning tasks should be retained for future use.

Lately, the concept of *lifelong learning* was defined in [23] as follows. Given a stream of n tasks $\{T_1, T_2, \ldots, T_n\}$ already learnt by a model, a knowledge base B can be obtained where all previously learnt knowledge is maintained. Note that, these tasks can be of different types and from different domains [23]. Then, when a new task T_{n+1} comes, the objectives of lifelong learning are mainly twofold: on the one hand, the existing knowledge in B should be leveraged to help optimise the performance of the new task T_{n+1}; on the other hand, the knowledge obtained from T_{n+1} should be integrated into B by updating it without causing a forgetting of prior knowledge of all past tasks.

From the above description, it could be noticed that the knowledge base B plays a crucial role throughout the whole learning process. This, to some extent, can gain insight from human learning mechanisms. We, as human beings, could learn better, easier, and faster, when we have more knowledge obtained on our previous life experiences [23]. Therefore, following the line of human intelligence, the crux of learning in a lifelong manner is to endow the system the capability of performing human-like knowledge-based learning.

However, in the context of deep learning, the *catastrophic forgetting* issue associated with the knowledge base B is considerably severe in most network systems. That is, when training a deep learning system (trained on some other task already) for a new task, the new learning process tends to interfere catastrophically with the previous learning. As a consequence, the performance of the system will be degraded heavily for past tasks, which is in contrast to the human brains and human learning systems [58]. To address this issue, in recent years, a number of lifelong learning techniques have been proposed and studied in the deep learning community, generally falling into three categories: dynamic architecture-based approaches, memory-based ones, and regularisation-based ones. For more details of various lifelong learning strategies, please refer to [23,24].

3.2. Elastic Weight Consolidation

In the current study, we focus on a regularisation-based lifelong learning approach called Elastic Weight Consolidation (EWC), which tackles the catastrophic forgetting problem by regularising the parameters in a network. The approach was first proposed by Kirkpatrick et al. [34] in 2017. Recently, it has been successfully exploited in several domains and applications [26,59]. Further, it has been derived into several relevant variants being investigated, e. g., EWC++ [60], online EWC [61], and R-EWC [62].

The idea of EWC is motivated by a consolidation process of human memory, which is known as synaptic consolidation or synaptic maintenance. This process enables us to consolidate previous memories within the related synapses to handle long-term memory tasks by reducing the plasticity of these synapses, and thus, these previously consolidated memories will not be altered by a future memory [63].

Similarly, in EWC, the plasticity of the task-relevant parameters (like the synapses in nervous systems) in a previously learnt model can be altered accordingly, to avoid changing significantly on these parameters when a future unseen task arrives. Especially, this is achieved by regularising the learning process with a quadratic penalty on the difference between the parameters for the prior and current tasks, the process of which is detailed in the following paragraphs.

In general, as aforementioned, given a prior optimised configuration of parameters $\theta^*_{1:n-1}$ for the past $n-1$ tasks, the objective of a lifelong learner is to learn an updated set $\theta^*_{1:n}$ for a new task T_n. In the following, an example when $n = 2$ is illustrated for an easy understanding, but EWC works when $n > 2$ as well. In this case, two datasets $\mathcal{D}_1 = (\mathcal{X}_1, \mathcal{Y}_1)$ and $\mathcal{D}_2 = (\mathcal{X}_2, \mathcal{Y}_2)$ are applied for the two tasks T_1 and T_2, respectively, with samples $x \in \mathcal{X}$ and their corresponding labels $y \in \mathcal{Y}$. Additionally, the sets of parameters Θ^*_1 and Θ^*_2 denote the configurations that deliver low loss (i. e., high performance) for θ_1 and θ_2, where θ_1 and θ_2 represent the parameters of T_1 and T_2, respectively.

In a conventional learning system where T_1 and T_2 are trained independently, our target is to find two configurations that meet the following criteria:

$$\theta^*_1 \in \Theta^*_1 \text{ and } \theta^*_2 \in \Theta^*_2, \tag{1}$$

where θ^*_1 and θ^*_2 denote the configurations of θ that result in a good performance for T_1 and T_2, respectively.

Nevertheless, in EWC, as T_2 is learnt after T_1, the optimisation of θ_2 is distinguished from that in Eq. (1), and can be given as follows:

$$\theta^*_2 \in \Theta^*_1 \cap \Theta^*_2, \tag{2}$$

where θ_2 is optimised with a constraint to stay in a low-error region of T_1 centred around θ^*_1. In this manner, a good configuration θ_2 should lie in the intersection of the low-error regions of T_1 and T_2 to prevent forgetting T_1, as depicted in Fig. 1.

Mathematically, when considering the learning process from a probabilistic perspective, we can model the relevance of the trainable parameters θ with respect to all the training data (i. e., $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$) as the posterior probability distribution $p(\theta|\mathcal{D})$. Then, the logarithm value of it can be decomposed by the Bayes' theorem:

$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}|\theta) + \log p(\theta) - \log p(\mathcal{D}). \tag{3}$$

Then, when splitting $\mathcal{D}$ into $\mathcal{D}_1$ and $\mathcal{D}_2$ and employing them one after the other, the above equation can be written as:

$$\begin{aligned}\log p(\theta|\mathcal{D}) = &\log p(\mathcal{D}_2|\mathcal{D}_1, \theta) + \log p(\theta|\mathcal{D}_1) \\ &- \log p(\mathcal{D}_2|\mathcal{D}_1),\end{aligned} \tag{4}$$

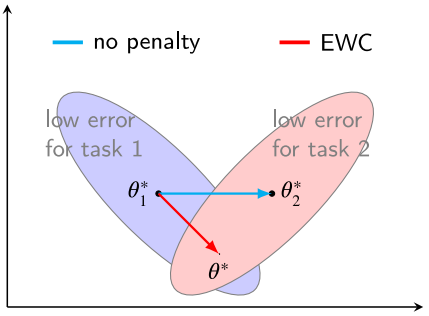


Fig. 1. Illustration of Elastic Weight Consolidation (EWC).

where $\mathcal{D}_1$ denotes the data for the prior task T_1, while $\mathcal{D}_2$ is for the current task T_2.

Furthermore, we assume that $\mathcal{D}_1$ and $\mathcal{D}_2$ are independent. In this circumstance, Eq. (4) can be reformulated as

$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}_2|\theta) + \log p(\theta|\mathcal{D}_1) - \log p(\mathcal{D}_2), \tag{5}$$

where the posterior probability term $\log p(\theta|\mathcal{D}_1)$ contains all the knowledge related to T_1. Hence, when implementing EWC, the aim is to get information about the parameter importance from $\log p(\theta|\mathcal{D}_1)$, and then take this information into account in the succeeding learning to prevent or at least mitigate the forgetting. Unfortunately, this posterior is intractable. To handle this issue, Laplace approximation is employed to approximate it as a Gaussian distribution with mean given by θ and a diagonal precision by the diagonal of the Fisher information matrix F as

$$F_i = \mathbb{E}_{\mathcal{D}_1}\left[\frac{\partial^2 \log p(\mathcal{D}_1|\theta)}{\partial \theta_i^2}\right]_{\theta=\theta^*_1}, \tag{6}$$

where i denotes the index of values on the diagonal of the matrix F. F_i then can be exploited to estimate the importance degree of each parameter to T_1. For instance, a great value of F_i indicates a high importance degree, and thus implies that it should not be changed much so that the performance of prior tasks can be maintained.

As a consequence, when training the network for T_2 after T_1, a penalty term with respect to T_1 is added to the objective function as

$$\mathcal{L}'(\theta) = \mathcal{L}(\theta) + \frac{1}{2}\lambda \cdot \sum_i F_i(\theta_i - \theta^*_{1,i})^2, \tag{7}$$

where $\mathcal{L}'(\theta)$ represents the new loss function in EWC, and $\mathcal{L}(\theta)$ is the loss for T_2 only. In addition, λ is a pre-defined hyperparameter to regulate how important the past tasks are compared to the current one, and i is the index of each trainable parameter.

In summary, when learning a new task, EWC penalises large changes to the most relevant parameters with respect to past tasks and suggests updating parameters for the new task mainly along directions with low Fisher information. In our case, this training strategy enables T_2 to be learnt without suffering catastrophic forgetting of T_1.

3.3. Continual emotion recognition

As discussed in Section 3.2, aiming at addressing catastrophic forgetting, a network can be learnt with the assistance of EWC, in which the importance of parameters with respect to prior tasks is exploited to selectively adjust the plasticity of parameters when a new task is given.

Although there are several studies that investigate this learning approach and its variants as aforementioned, we believe that

Algorithm 1: The training process of EWC-based continual cross-cultural emotion recognition.

Initialise:
N: the total number of databases;
n: index to indicate the nth database;
$\mathcal{D}_n = (\mathcal{X}_n, \mathcal{Y}_n)$: the nth database;
θ: parameters of the model;
θ_0: initialised parameters of θ;
for $n = 1, \ldots, N$ **do**
  if $n = 1$ **then**
    optimise θ via minimising a conventional loss;
  else
    compute the Fisher Information for the prior $\mathcal{D}_{n-1}$;
    save the prior configuration θ^*_{n-1};
    optimise θ via minimising an EWC-based loss function;
  end
  $\theta \leftarrow \theta^*_n$;
end

this is the first attempt to explore it in the field of audiovisual affective computing. In particular, in this work, cross-cultural emotion recognition is taken as a paradigm to investigate its effectiveness and efficiency. Being an active research area in affective computing nowadays, the main goal of cross-cultural emotion recognition is to establish a compatible and efficient framework to handle the discrepancy across multiple cultures, providing a good test bed of the proposed continual affective computing.

In this case, a neural network will be learnt to estimate emotion patterns, via learning sequentially from a series of databases $\{\mathcal{D}_n\}$ with $n = 1, 2, \ldots, N$, each consisting of plenty of emotional instances from one specific culture. After training on one database for the current culture T_n, the Fisher information can be estimated as given in Eq. (6), and in the meanwhile, the best configuration of θ^*_n is kept for future usage. Then, when given a new database of another culture T_{n+1}, the network will be optimised using Eq. (7). These procedures can be performed repeatedly until all N cultures have been used for training the network. The pseudo-code describing these procedures is also presented in Algorithm 1.

Our expectation is that after training the Nth culture, the performance of all prior $N - 1$ tasks is not heavily degraded.

4. Experimental implementation

To evaluate the effectiveness and efficiency of the proposed continual emotion learning in a cross-culture scenario, extensive experiments were carried out. In this section, the databases and features applied in our experiments are first given in Section 4.1. Then, we detail the experimental setups for the sake of experiment replication in Section 4.2. After that, we describe the evaluation measures for the performance comparison in Section 4.3.

Before going into each part, we would like to introduce the international Audio/Visual Emotion Challenge (AVEC), from which the employed corpora and experimental settings in this work are originated. This challenge has been organised from 2011 to 2019, and aims to automatically and accurately detect subjects' emotional states continuously through the acoustic, visual, and also physiological signals [30,64]. In each year, two to three related databases are selected and benchmarked for the challenge.

Table 1
Statistical information of RECOLA and SEWA datasets over training, development, and test partitions.

	RECOLA		SEWA	
	# users	# segments	# users	# segments
train	9	67 500	34	55 072
dev	9	67 500	14	22 307
test	9	67 500	16	27 597
$\sum$	27	202 500	64	104 976

4.1. Evaluated databases and features

For the given task, two multimodal emotional datasets were chosen, i. e., the RECOLA dataset and the SEWA database. In the following, we briefly introduce these two datasets and the selected features.

4.1.1. RECOLA & SEWA

The Remote Collaborative and Affective Interactions dataset (RECOLA) [65], is a standard database used in the AVEC challenge series since 2015 [30]. This database contains audiovisual recordings of spontaneous and natural interactions from 27 French-speaking participants in order to investigate socio-affective behaviours in the context of remote collaborative tasks [65]. In particular, time- and value-continuous dimensional annotations in terms of arousal and valence are provided with a constant frame rate of 40 ms for the first five minutes of each recording [65]. For the aim of the challenge, the dataset was then further divided into three disjoint partitions (i. e., training, development, and test), by well balancing the age and gender of the participants. Consequently, each partition includes nine recordings, resulting in 67 500 segments in total for each partition. For full details on the RECOLA database, please refer to [65].

The Automatic Sentiment Analysis in the Wild database (SEWA) [29], is a multicultural corpus consisting of audiovisual recordings of human interactive behaviours in naturalistic contexts from six different cultural backgrounds. Specifically, the recordings from German participants have been exploited in the AVEC challenge series since 2017 [30]. Following the advice shown in [30], our experiments are performed on the 64 German recordings. Moreover, time- and value-continuous dimensional affect ratings with respect to arousal and valence are available at a constant frame rate of 100 ms. Similar to RECOLA, the full German corpus was further split into three subject-independent parts, i. e., 34 recordings for training, 14 ones for development, and the remaining 16 ones for testing. As a result, 55 072, 22 307, and 27 597 annotated segments are obtained for the training, development, and test set, respectively. For full details on this database, please refer to [29].

Overall, the general statistical information of the aforementioned datasets RECOLA and SEWA is shown in Table 1.

4.1.2. Features

To extract features from these databases, bag-of-audio-words (BoAW) representations and bag-of-video-words (BoVW) features were generated with the help of our open-source toolkit openXBOW [66]. Note that, these features are provided in the AVEC challenge recently as established feature sets to develop appropriate baselines. Therefore, we employed the same feature sets for a fair performance comparison with other related works.

In particular, for audio data, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS [67]) was extracted as Low-Level Descriptors (LLDs) for each frame via our openSMILE toolkit [68]. Then, BoAW representations were computed over a collection of successive frames for each step of 40 ms and 100 ms to match the frequency of the annotations of RECOLA and SEWA,

respectively. Note that, we followed the settings as provided in the AVEC challenge baseline systems [30], i. e., segment-level representations were computed from the LLDs with one hard assignment on a codebook of size 100. This resulted in a set of 100 acoustic BoAW features.

Similarly, following the suggestions in [30], 17 Facial Action Units (FAUs) were first extracted per frame as LLDs via the open-source toolkit OpenFace [69]. Then, we applied the same processing chain as when generating BoAW features. With this process, we got additional 100-dimensional visual representations, on both the RECOLA and SEWA sets accordingly.

4.2. Experimental setup

The proposed EWC-based continual emotion recognition method was evaluated on the aforementioned datasets. Note that, for a better understanding and a clearer view, we defined the training with the RECOLA database as FR, and the learning on the SEWA German database as GE. This led to the following two distinct sequential training schemes:

- *FR after GE, w/ EWC*, where the model is first trained on SEWA, and later trained on RECOLA, with the EWC constraint;
- *GE after FR, w/ EWC*, where RECOLA is first employed to train the model, and SEWA as the new task to be learnt.

Additionally, for comparison, we carried out other baselines as well. To be more specific, the following baseline systems were run:

- single task only, where we performed isolated learning on each dataset separately. In this scheme, we had two baselines which are marked as *FR only* and *GE only*;
- single task only with weight regularisation, where the model was again optimised on a single dataset. However, the L2 regularisation penalty term was explored for better generalisation performance. With this scenario, another two baselines were provided, i. e., *FR only, w/ L2-norm* and *GE only, w/ L2-norm*;
- furthermore, we considered sequential fine-tuning as an extra baseline, where the model is fine-tuned on the second corpus after firstly having been optimised via the first one. This is similar to the proposed training process, however, without considering any constraint. In this case, we have *FR after GE, w/o EWC* and *GE after FR, w/o EWC*.

Besides of these outlined baseline systems, we also ran joint training, i. e., training the network on all available datasets jointly, denoted hereinafter as *FR and GE*. The hypothesis is that the introduced lifelong learning based sequential training scheme could perform better than the isolated learning scheme (single task only) and the fine-tuning based sequential training scheme, and meanwhile it could perform closely, or even competitively, to the joint training scheme.

Moreover, within each training scheme, models were individually trained on acoustic features, visual features, or the combination of the two, for arousal and valence prediction, respectively. To achieve this aim, we implemented all models by using deep recurrent neural networks with gated recurrent units (GRUs). GRUs are frequently applied in replacement of long short-term memory units and can deliver comparable performance [70]. Note that, for the sake of fair comparison, the same network structure was used for all training schemes, and the network settings were empirically chosen which can provide competitive performance on both databases when comparing with other previous within-cultural models [30]. To be more specific, each network consists of four hidden layers with 100 units per layer. While training the network, the Adam optimiser was employed with an initial learning rate of 0.001.

In addition, the early stopping strategy was executed if no performance improvement on the development set was observed after 20 successive epochs. It is important to note that, the development set shared the same culture type with the training set of the current task. For instance, when training a model following *FR after GE* (learn SEWA first, then RECOLA), the learning process first is ceased based on the performance on the SEWA development set when learning on the SEWA training set. After that, when continually learning on the RECOLA training set as a new task, the training process will be terminated by inspecting its performance on the RECOLA development set.

Furthermore, in all of the experiments, following the suggestions of the AVEC challenges [30], annotation shifting was performed to compensate annotation delay, and a post-processing chain of four stages was performed with an aim to refine the obtained predictions. For more details, please refer to [30].

4.3. Evaluation metrics

To evaluate the performance of the EWC-based continual emotion regression system, we computed the *Concordance Correlation Coefficient* (CCC), which was officially utilised in the AVEC challenge series [30]. Formally, the CCC is defined by:

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{8}$$

where ρ denotes the Pearson's correlation coefficient (PCC) between two time-series (i. e., our predictions and the gold-standards in our case), μ_x and μ_y stand for the mean of each time series, and σ_x^2 and σ_y^2 represent the corresponding variances. CCC is preferred over PCC as it considers not only the shape similarity between the two series but also the bias term $(\mu_x - \mu_y)^2$. This is especially relevant for evaluating the performance of time-continuous emotion prediction models, as both the trends as well as the absolute prediction values are vital. The value of *CCC* is within the range of $[-1, 1]$, where $+1$ represents perfect concordance, -1 total discordance, and 0 no concordance at all. Hence, a higher *CCC* implies better system performance.

5. Experimental results and discussions

In this section, we present and discuss the prediction results obtained with the proposed method. In particular, we compare our models against other baseline systems. In addition, we further investigate the effectiveness of the method by visualising the effect of the elastic-weight-based penalty term and its impact on model parameters.

5.1. Cross-cultural emotion recognition

Table 2 demonstrates the results of various training strategies for arousal and valence predictions on the RECOLA and SEWA datasets, respectively. For a clear view, we sort these training strategies into three categories, i. e., six baseline strategies to compare against, two lifelong training models, and a joint training strategy which can be viewed as an upper bound in our case. Moreover, results are summarised in three blocks with respect to the applied feature sets, namely, audio, video, and the combination of the two.

First, we compare our models against six baselines. From the table, one may notice that the performance is heavily degraded when there is a cultural mismatch between training and inferring

Table 2
CCC performances via various training strategies for emotion regression based on audio, video, or their combination. Performance on the development sets and test sets of the two databases (FR_{dev}, FR_{test}, GE_{dev}, GE_{test}) as well as the average performance on the two test sets (μ_{test}) are reported for arousal and valence, respectively.

Features	Methods	Arousal					Valence				
		FR_{dev}	FR_{test}	GE_{dev}	GE_{test}	μ_{test}	FR_{dev}	FR_{test}	GE_{dev}	GE_{test}	μ_{test}
Audio	Baseline strategies										
	FR only	.631	.552	.311	.036	.294	.287	.248	.085	.024	.136
	GE only	−.022	.009	.388	.246	.128	.007	−.003	.390	.245	.121
	FR only, w/ *L2*-norm	.613	.543	.276	.049	.296	.227	.249	.140	.019	.134
	GE only, w/ *L2*-norm	−.023	.014	.357	.224	.119	.025	.046	.361	.240	.143
	FR after GE, w/o EWC	.640	.538	.334	.047	.293	.325	.272	.072	−.011	.131
	GE after FR, w/o EWC	−.057	−.006	.383	.243	.119	.027	.050	.360	.251	.151
	Proposed lifelong learning strategies										
	FR after GE, w/ EWC	.639	.538	.335	.046	.292	.232	.159	.265	.156	.158
	GE after FR, w/ EWC	.554	.509	.377	.157	.333	.019	-.009	.406	.192	.092
	Joint training (upper bound)										
	FR and GE	.498	.457	.389	.176	.317	.293	.234	.343	.186	.210
Video	Baseline strategies										
	FR only	.167	.180	.277	.175	.178	.413	.354	.552	.397	.376
	GE only	.073	.014	.517	.374	.194	.406	.211	.571	.517	.364
	FR only, w/ *L2*-norm	.181	.217	.223	.227	.222	.432	.341	.561	.413	.377
	GE only, w/ *L2*-norm	.094	.029	.563	.404	.217	.396	.215	.581	.508	.362
	FR after GE, w/o EWC	.195	.241	.229	.212	.227	.430	.355	.480	.289	.322
	GE after FR, w/o EWC	.078	.024	.536	.365	.195	.368	.197	.564	.467	.332
	Proposed lifelong learning strategies										
	FR after GE, w/ EWC	.195	.241	.230	.212	.227	.479	.370	.583	.465	.418
	GE after FR, w/ EWC	.197	.184	.601	.413	.299	.466	.271	.562	.587	.429
	Joint training (upper bound)										
	FR and GE	.168	.177	.541	.455	.316	.550	.382	.598	.600	.491
Fusion	Baseline strategies										
	FR only	.621	.578	.397	.102	.340	.529	.436	.334	.189	.313
	GE only	.090	.063	.581	.401	.232	.241	.134	.623	.485	.310
	FR only, w/ *L2*-norm	.634	.627	.351	.145	.386	.511	.380	.372	.222	.301
	GE only, w/ *L2*-norm	.080	.050	.588	.412	.231	.249	.134	.633	.536	.335
	FR after GE, w/o EWC	.631	.596	.436	.150	.373	.528	.421	.381	.250	.336
	GE after FR, w/o EWC	.056	.035	.613	.424	.230	.250	.150	.637	.527	.339
	Proposed lifelong learning strategies										
	FR after GE, w/ EWC	.600	.599	.500	.224	.412	.514	.412	.534	.370	.391
	GE after FR, w/ EWC	.551	.530	.568	.366	.448	.450	.252	.617	.569	.411
	Joint training (upper bound)										
	FR and GE	.534	.533	.601	.399	.466	.523	.399	.601	.584	.492

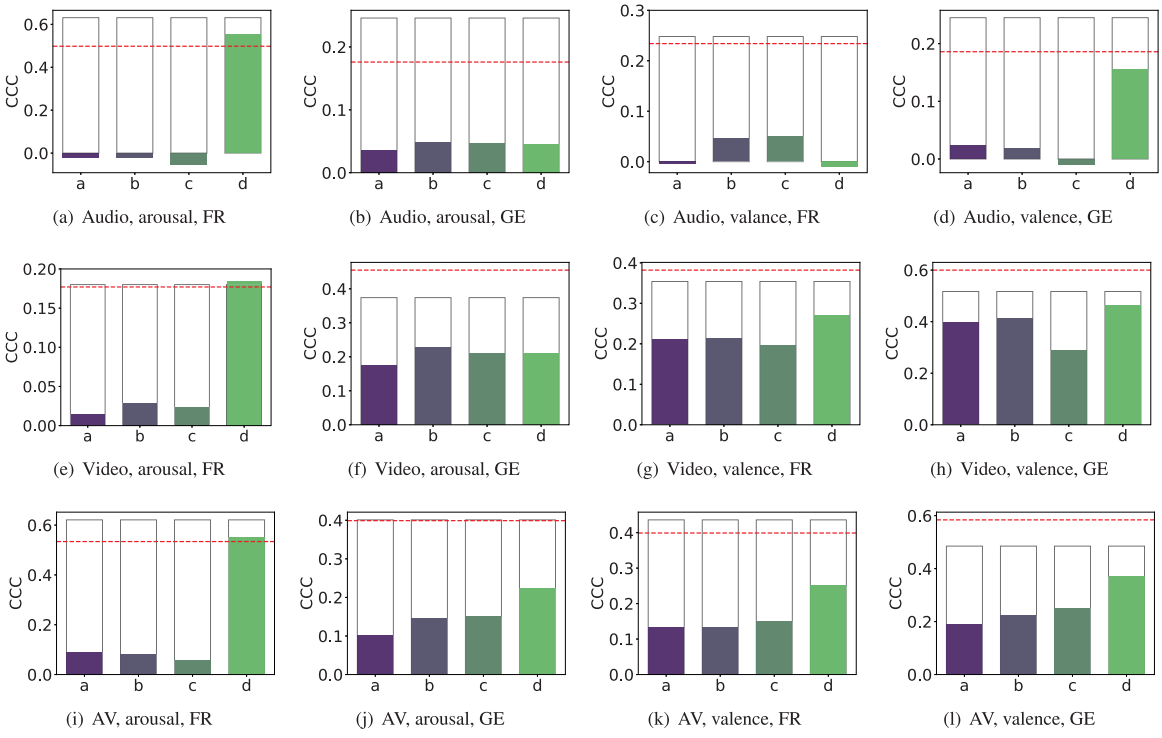
sets. We may take the arousal prediction from audio as an example. On FR_{test}, a CCC of .552 is obtained by training on the same cultural data (i.e., *FR only*), while the performance dramatically reduces to .009 if training with German data only (i.e., *GE only*). Likewise, for GE_{test}, the performance of *GE only* is remarkably superior to *FR only*. Similar observations can be drawn over all three distinct feature sets for the arousal prediction as well as for the valence prediction.

Moreover, results of another two baseline training models, *FR only, w/ L2-norm* and *GE only, w/ L2-norm*, are also provided in Table 2. These two models aim at improving the generalisation performance on new, unseen data. For instance, when comparing *FR only* and *FR only, w/ L2-norm*, the obtained CCCs on both test sets are boosted for arousal via the fused audiovisual features, from .578 to .627 on FR_{test} and from .102 to .145 on GE_{test}. Nevertheless, a severe performance discrepancy still exists between the two sets, and this indicates that it is essential to construct a model to learn from data of both cultures.

Further, in order to learn from data of both sets, sequential training strategies have been evaluated in the last two baseline systems, namely *FR after GE, w/o EWC* and GE after FR, w/o EWC. These systems, without considering the EWC regularisation, suffer severely from the catastrophic forgetting issue. Let us take arousal prediction from audio signals with *GE after FR, w/o EWC* as an example, where the model first learns from *FR* and then *GE*. The obtained CCC for French decreases dramatically, from .631 to −.057 on FR_{dev} and from .552 to −.006 on FR_{test}. It can be seen that, though both tasks are learnt in a sequence, performance of the first task is damaged as the model adaptation to the second culture disrupts the knowledge learnt from the first one. It suggests that advanced training strategies are inevitable and essential to deal with this issue.

With the proposed continual learning approaches, one may notice that the aforementioned catastrophic forgetting problem is alleviated remarkably, by preserving the knowledge of previous tasks via the EWC regularisation during training. Again, taking for instance predicting arousal from audio signals, with EWC, the CCCs achieved on the French dataset by *GE after FR, w/ EWC* are .554 on FR_{dev} and .509 on FR_{test}, respectively, and in the meanwhile the CCCs for German are still competitive to a German-dependent model (.377 vs .388 on GE_{dev} and .157 vs .246 on GE_{test}). Such an observation can be found in Table 2 in other scenarios.

Notably, when comparing *FR after GE* with *GE after FR*, it is also interesting to observe that, the performance of the latter is on average superior to the former. For instance, when comparing the average performance on the two test sets μ_{test}, *GE after FR, w/ EWC* is better than *FR after GE, w/ EWC* in five out of six cases (three feature sets by two emotion dimensions) except for the valence prediction from audio signals. This may indicate



(a) Audio, arousal, FR (b) Audio, arousal, GE (c) Audio, valance, FR (d) Audio, valence, GE

(e) Video, arousal, FR (f) Video, arousal, GE (g) Video, valence, FR (h) Video, valence, GE

(i) AV, arousal, FR (j) AV, arousal, GE (k) AV, valence, FR (l) AV, valence, GE

Fig. 2. Visualisation of the performances in terms of CCC of the proposed methods comparing with other baseline approaches on the test sets of FR and GE. Results are separately shown for arousal and valence regressions via audio, video, and their combination (AV), respectively. Note that, the white bars indicate the performance of a matched culture-specific model, while the red dotted lines denote the performance of a joint training model. a: mismatched culture-specific model, b: mismatched model w/ L2-norm, c: sequential training w/o EWC, d: sequential training w/ EWC (ours).

that the order of the training tasks also plays a key role in a continual emotion recognition system. We can gain some insight from curriculum learning and infer that learning several tasks in a proper order might lead to improved average performance.

Furthermore, from Table 2 one may observe that, in most cases, when modelling emotion patterns from audio only, the models achieve better performance in the arousal prediction than in the valence prediction. In contrast, when estimating via facial expressions, observations are found in another way around. Moreover, when combining audio and visual features via early fusion, the performance is improved. These findings are consistent with previous studies [71,72]. In particular, when learning from both audio and visual signals, the best average CCCs on the two test sets μ_{test} are obtained by joint training (*FR and GE*), reaching to .466 for arousal and .492 for valence, respectively.

Rather than the joint training paradigm, in our proposed models, the two datasets were learnt one after the other, and this reduces the high storage requirement issue we might face in joint training. With this manner, comparable average performance μ_{test} in terms of CCC is achieved with our EWC-based model *GE after FR, w/ EWC*, i. e., .448 and .411 for arousal and valence, respectively. This suggests that sequential learning is, to some degree, a potential replacement of joint training as the system may benefit from lower storage requirement and computation load. This is extremely vital for real-life intelligent systems, where the number of given tasks and thus the amount of training data might grow rapidly.

For a better interpretation, these results are also visualised in Fig. 2. In this figure, the performance of the proposed systems on FR_{test} and GE_{test} are compared with their corresponding baseline systems under six distinct setting combinations of three different feature sets and two emotion dimensions separately. Moreover, the performance of two upper bound systems is depicted as well. In particular, in each subfigure, the upper bound of its matched culture-specific model (e. g., *FR only*) is presented as a white bar; while the joint training upper bound is drawn as red dotted lines. Another four coloured bars denote the corresponding performance of the mismatched culture-specific model (e. g., *GE only*), the mismatched model with L2-norm (e. g., *GE only, w/ L2-norm*), the sequential training model without EWC (e. g., *GE after FR, w/o EWC*), and the introduced sequential training model with EWC (e. g., *GE after FR, w/ EWC*, also the lifelong learning model), respectively. As a consequence, the performance degradation from all introduced and compared models can be visualised as the white space between two bars.

We can see from Fig. 2 that, in most cases, our lifelong learning model outperforms other baseline models and yields less performance difference with its matched culture-specific model than other baseline models. In particular, when training for arousal, our *GE after FR, w/ EWC* models achieve competitive or even better results than the joint training upper bounds (cf. Fig. 2(a), (e), and (i)). This demonstrates the great potential of implementing continual emotion recognition by investigating advanced lifelong learning algorithms.

5.2. Hyperparameter selection

As given in Eq. (7), λ is utilised to regulate the contribution of previous knowledge when learning a new task. Therefore, to better illustrate the effect of λ for the performance, we trained various models with different λs in the range of $[10^{-2}, 10^{10}]$. Fig. 3 shows how the performance of our lifelong model varies with respect to various λs on the development sets. Note that, in this figure, only results on *GE after FR, w/ EWC* models are demonstrated, as it generally performs better than *FR after GE, w/ EWC* (cf. Section 5.1).

As shown in Fig. 3, when λ is small, performance on the previous task (i. e., French emotion recognition) is relatively low, compared with the second task (i. e., German emotion recognition). Then, when λ increases, the performance on FR_{dev} improves

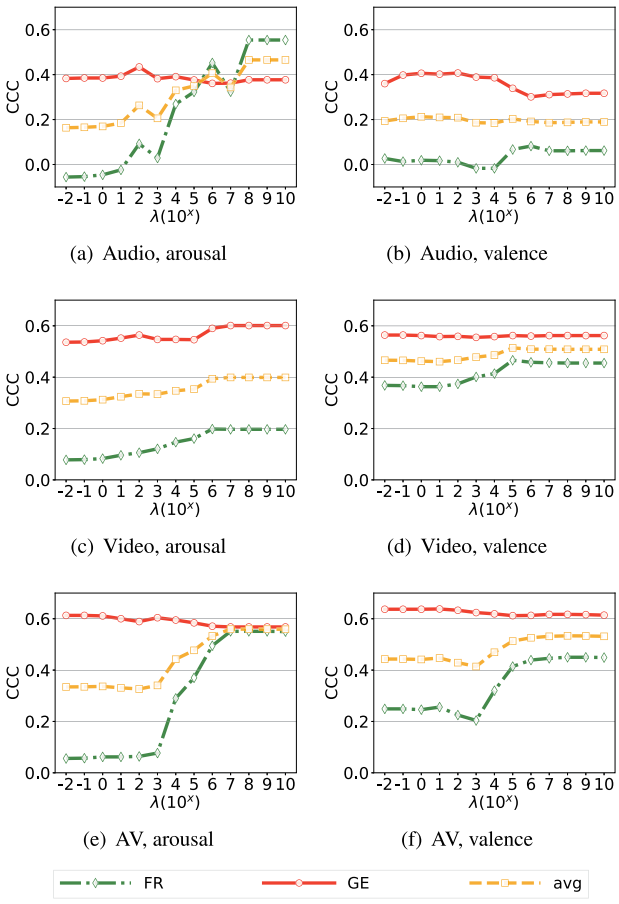


Fig. 3. The effect of the hyperparameter to control the regularisation λ in the proposed *GE after FR, w/ EWC* model, when predicting arousal and valence on the development sets via three various feature types, i. e., audio, video, and audiovisual (AV). The average performance of both FR and GE is calculated and denoted as avg.

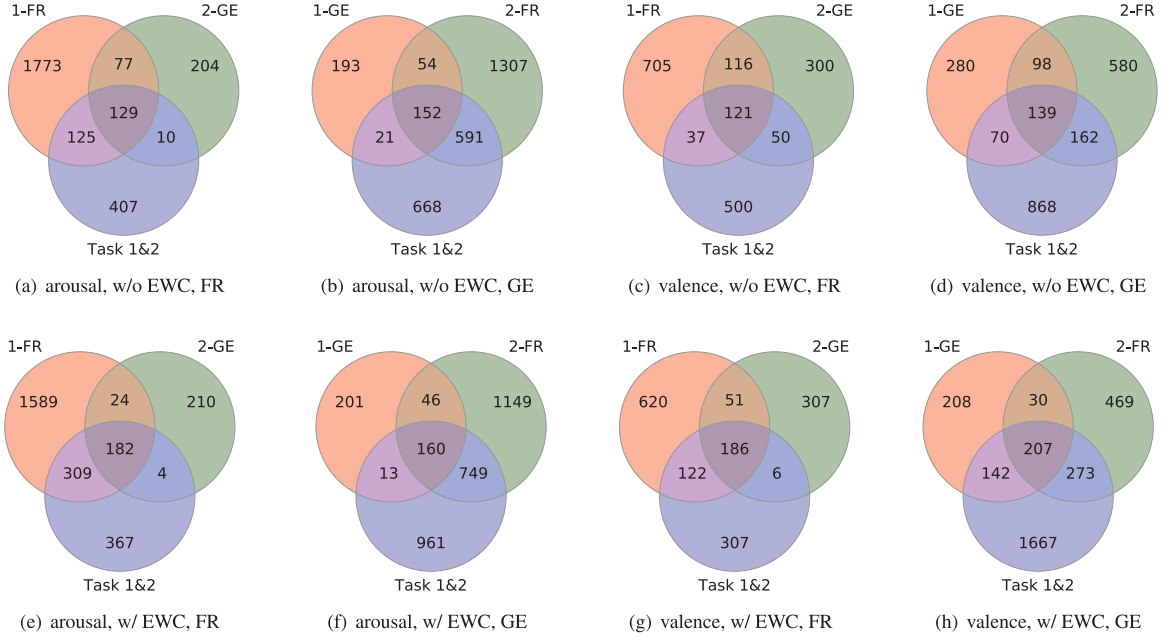


Fig. 4. Venn diagrams to visualise the relations among three parameter sets by analysing the important parameters obtained in three models for audiovisual emotion prediction, where each of them can be viewed as a circle. In particular, the red circle denotes a culture-specific model, i. e., 1-FR or 1-GE; the green circle represents another culture-specific model, i. e., 2-GE or 2-FR; and the purple circle indicates a sequential training model that learns task 1 and 2 sequentially. The values in the circles show the number of important parameters, which belong to one set only or lie at the intersection of two or even three sets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in a large margin in most cases, until a point where it does not benefit from further increasing. Likewise, when depicting the average performance on FR_{dev} and GE_{dev}, a performance improvement can also be observed along the increased λ until it becomes flat again. In this regard, a proper λ is demanded to retain the previous knowledge. However, the performance on GE_{dev} either increases (cf. Fig. 3(c)), remains (cf. Fig. 3(a) and (d)), or decreases slightly (cf. Fig. 3(b), (e), and (f)) under different settings. This may highly depend on whether the previous knowledge is beneficial or not to the current learning process.

5.3. Effectiveness verification

In the following, we further verify the effectiveness of the EWC-based sequential learning approach by inspecting the impacts of EWC on the plasticity of parameters to be learnt. In particular, in our selected RNN model for emotion regression tasks, there are more than 271 K parameters to be learnt. Of these parameters, their importance with respect to a given task can be estimated by its Fisher information (cf. Section 3.2). On this account, given a predefined threshold 10^{-4} to only consider the parameters that have a Fisher value above it, for each model, a parameter set can be generated to incorporate these important parameters. Then, when EWC is applied, the plasticity of the parameters in the set can be decreased to tackle catastrophic forgetting when a new task comes.

In this study, we look into the relations of three parameter sets of this kind, i. e., one for a model trained on French only, one for German only, and one for a model trained on the two databases one after another. Then, the relations of these three sets can be demonstrated by a couple of Venn diagrams, under eight audiovisual training scenarios, as shown in Fig. 4. In particular, the four Venn diagrams in the upper row of Fig. 4 present the four cases when carrying out sequential learning without EWC, while the remaining four in the lower row are corresponding models trained with EWC. Moreover, the values in the diagram depict the number of parameters which are vital to only one model, or two models (overlaps of every two circles), or three models (overlaps of all three circles). Hence, when investigating the intersection of a red circle (model trained only on the first task) and a purple circle (model learnt on two tasks) and comparing every two Venn diagrams in a column, one may see that the intersection in the lower diagram is always greater than (three out of four cases) or at least equal to (one case only) that from the upper diagram. This might indicate that sequential training with EWC is capable of maintaining more parameters that are important to the previous task, by reducing the plasticity of these parameters for future learning.

5.4. Discussion

Based on our previous results, it has been shown that, by quantifying the importance of weights to previous tasks and then adjusting the plasticity of weights accordingly for new tasks, our EWC-based continual emotion recognition models outperform other baseline models. The proposed method overcomes the limitations of the conventional isolated training approach, and becomes a promising alternative of joint training to dealing with the discrepancy among multiple cultures for emotion perception.

To close this section, we would like to point out some potential limitations of the current EWC-based continual emotion recognition system. As can be seen in Fig. 4, after learning two tasks, the number of important parameters is increasing to preserve the knowledge from both tasks, indicating that a lower amount of parameters is able to be largely changed for future tasks. Due to this limitation, EWC is not sufficient for learning a large number of tasks, when most, if not all, parameters appear to be rather essential to keep the knowledge for all past tasks. The reason is that it might be caught in a dilemma: on the one hand, new knowledge can only be acquired by updating parameters accordingly; on the other hand, past knowledge will get destroyed if any parameter is modified.

Moreover, although the performance of the proposed model is superior to other baseline models, meaning that the catastrophic forgetting problem is partially addressed, much work is still needed towards closing the performance gap between it and a culture-specific model. Therefore, in future studies, other more advanced lifelong learning approaches will be investigated and applied towards general emotion perception systems. These techniques include, but are not limited to, PathNet [73], GeppNet [74], progressive neural networks [75], and dynamically expandable networks [76].

Finally, as the first work towards lifelong learning in audiovisual affective computing, merely two datasets cross languages and cultures were considered. In the future, we plan to investigate the generalisation and robustness of the proposed lifelong learning paradigm cross multiple corpora like the ones used in [77], or cross multiple tasks. This work is of importance as it leads to a more realistic application scenario.

6. Conclusion

To prevent the catastrophic forgetting problem of conventional machine learning algorithms, in this article we presented a lifelong learning perspective for affective computing. More specifically, we proposed the lifelong learning algorithm of Elastic Weight Consolidation (EWC) in a well-established application of cross-cultural audiovisual emotion recognition. Through comprehensive experiments and analysis, it is found that EWC enables a model to learn multiple tasks in an open-set environment, with no or limited forgetting about the knowledge obtained from previous tasks. These findings will facilitate the development and implementation of affective computing into a real-life scenario, because of its heterogeneous and dynamic characteristics.

In future work, we will further investigate other advanced lifelong learning algorithms, and compare their performance with the introduced EWC in the context of affective computing. Moreover, towards accomplishing emotional intelligence in the context of Internet of People (IoP), we are interested in investigating the proposed emotion recognition model to support various applications, such as monitoring health states in smart homes [78, 79], reducing energy consumption according to personal living habit in smart workplaces [80,81], and detecting negative emotions of group in smart cities [82,83].

Furthermore, in this research work, we focused on learning emotion perception models based on the audio and visual observations. Other than these two observations, additional factors such as the pressure and preferences of the users may also have an impact on both how they perceive others' emotions and how they express their own emotions. Hence, in the future, it is of great interest to take these subjective factors into account, for simulating the complexity of emotional experience, in the context of IoP.

CRediT authorship contribution statement

Jing Han: Conceptualization, Methodology, Writing - original draft, Software. **Zixing Zhang:** Conceptualization, Methodology, Writing - original draft. **Maja Pantic:** Writing - review & editing, Supervision. **Björn Schuller:** Writing - review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the TransAtlantic Platform "Digging into Data" collaboration grant (ACLEW: Analysing Child Language Experiences Around The World), with the support of the UK's Economic & Social Research Council through the research Grant No. HJ-253479. We also would like to thank our colleague Maximilian Schmitt for his help with the BoVW feature extraction of the RECOLA database.

References

[1] J. Miranda, N. Mäkitalo, J. Garcia-Alonso, J. Berrocal, T. Mikkonen, C. Canal, J.M. Murillo, From the Internet of Things to the Internet of People, IEEE Internet Comput. 19 (2) (2015) 40–47.
[2] M. Conti, A. Passarella, S.K. Das, The internet of people (IoP): A new wave in pervasive mobile computing, Pervasive Mob. Comput. 41 (2017) 1–27.
[3] J. Gubbi, R. Buyya, S. Marusic, M. Palaniswami, Internet of Things (IoT): A vision, architectural elements, and future directions, Future Gener. Comput. Syst. 29 (7) (2013) 1645–1660.
[4] J.S. Silva, P. Zhang, T. Pering, F. Boavida, T. Hara, N.C. Liebau, People-centric Internet of Things, IEEE Commun. Mag. 55 (2) (2017) 18–19.
[5] S. Chen, T. Liu, F. Gao, J. Ji, Z. Xu, B. Qian, H. Wu, X. Guan, Butler, not servant: A human-centric smart home energy management system, IEEE Commun. Mag. 55 (2) (2017) 27–33.
[6] U. Lopez-Novoa, U. Aguilera, M. Emaldi, D. López-De-Ipiña, I. Pérez-De-Albeniz, D. Valerdi, I. Iturricha, E. Arza, Overcrowding detection in indoor events using scalable technologies, Pers. Ubiquitous Comput. 21 (3) (2017) 507–519.
[7] M. Qiu, K. Gai, B. Thuraisingham, L. Tao, H. Zhao, Proactive user-centric secure data scheme using attribute-based semantic access controls for mobile clouds in financial industry, Future Gener. Comput. Syst. 80 (2018) 421–429.
[8] B. Farahani, F. Firouzi, V. Chang, M. Badaroglu, N. Constant, K. Mankodiya, Towards fog-driven IoT eHealth: Promises and challenges of IoT in medicine and healthcare, Future Gener. Comput. Syst. 78 (2018) 659–676.
[9] D. Casado-Mansilla, P. Garaizar, D. López-de Ipiña, User involvement matters: The side-effects of automated smart objects in pro-environmental behaviour, in: Proc. 9th International Conference on the Internet of Things, Bilbao, Spain, 2019, pp. 1–4.
[10] R.W. Picard, Affective Computing, MIT press, Cambridge, MA, 1997.
[11] J.A. Russell, Core affect and the psychological construction of emotion, Psychol. Rev. 110 (1) (2003) 145–172.
[12] A. Beatty, Anthropology and emotion, J. R. Anthropol. Inst. 20 (3) (2014) 545–563.
[13] B. Schuller, A. Batliner, Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing, John Wiley & Sons, Hoboken, NJ, 2013.
[14] P. Li, Y. Song, I. McLoughlin, W. Guo, L. Dai, An attention pooling based representation learning method for speech emotion recognition, in: Proc. INTERSPEECH, Hyderabad, India, 2018, pp. 3087–3091.
[15] J. Han, Z. Zhang, Z. Ren, F. Ringeval, B. Schuller, Towards conditional adversarial training for predicting emotions from speech, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Calgary, Canada, 2018, pp. 6822–6826.
[16] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, F. Hutter, Efficient and robust automated machine learning, in: Proc. Advances in Neural Information Processing Systems, NIPS, Montréal, Canada, 2015, pp. 2962–2970.
[17] Q. Yao, M. Wang, Y.-Q. Hu, Y.-F. Li, W.-W. Tu, Q. Yang, Y. Yu, Taking human out of learning applications: A survey on automated machine learning, 2018, arXiv preprint arXiv:1810.13306.
[18] Z. Zhang, N. Cummins, B. Schuller, Advanced data exploitation for speech analysis – An overview, IEEE Signal Process. Mag. 34 (4) (2017) 107–129.
[19] F. Eyben, M. Wöllmer, B. Schuller, A multitask approach to continuous five-dimensional affect sensing in natural speech, ACM Trans. Interact. Intell. Syst. 2 (1) (2012) 1–29.
[20] S.A. Taylor, N. Jaques, E. Nosakhare, A. Sano, R. Picard, Personalized multitask learning for predicting tomorrow's mood, stress, and health, IEEE Trans. Affect. Comput. (2017) 14.
[21] B. Zhang, E.M. Provost, G. Essl, Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences, IEEE Trans. Affect. Comput. 10 (1) (2019) 85–99.
[22] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[23] Z. Chen, B. Liu, Lifelong Machine Learning, Morgan & Claypool, San Rafael, CA, 2018.
[24] G.I. Parisi, R. Kemker, J.L. Part, C. Kanan, S. Wermter, Continual lifelong learning with neural networks: A review, Neural Netw. 113 (2019) 54–71.
[25] S. Thrun, T.M. Mitchell, Lifelong robot learning, Robot. Auton. Syst. 15 (1–2) (1995) 25–46.
[26] Z. Chen, N. Ma, B. Liu, Lifelong learning for sentiment classification, in: Proc. the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL, Beijing, China, 2015, pp. 750–756.
[27] Q. Ha, B. Nguyen-Hoang, M. Nghiem, Lifelong learning for cross-domain Vietnamese sentiment classification, in: Proc. 5th International Conference on Computational Social Networks, CSoNet, Ho Chi Minh City, Vietnam, 2016, pp. 298–308.
[28] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2008) 39–58.
[29] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, B. Schuller, K. Star, E. Hajiyev, M. Pantic, SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 18 pages.
[30] F. Ringeval, B. Schuller, M. Valstar, et al., AVEC 2019 workshop and challenge: State-of-Mind, depression with AI, and cross-cultural affect recognition, in: Proc. 9th AudioVisual Emotion Challenge, AVEC Associated with ACM Multimedia, Nice, France, 2019, p. 10.
[31] S. Hareli, K. Kafetsios, U. Hess, A cross-cultural study on emotion expression and the learning of social norms, Front. Psychol. 6 (2015) 1501.
[32] N. Lim, Cultural differences in emotion: differences in emotional arousal level between the East and the West, Integr. Med. Res. 5 (2) (2016) 105–109.
[33] R. Srinivasan, A.M. Martinez, Cross-cultural and cultural-specific production and perception of facial expressions of emotion in the wild, IEEE Trans. Affect. Comput. (2018) 15 pages.
[34] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A.A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al., Overcoming catastrophic forgetting in neural networks, Proc. Natl. Acad. Sci. 114 (13) (2017) 3521–3526.
[35] P. Song, W. Zheng, Feature selection based transfer subspace learning for speech emotion recognition, IEEE Trans. Affect. Comput. (2018) 11 pages.
[36] S. Albanie, A. Nagrani, A. Vedaldi, A. Zisserman, Emotion recognition in speech using cross-modal transfer in the wild, in: Proc. ACM International Conference on Multimedia, MM, Seoul, Korea 2018, pp. 292–301.
[37] J. Han, Z. Zhang, Z. Ren, B. Schuller, EmoBed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings, IEEE Trans. Affect. Comput. (2019) 12.
[38] X. Glorot, A. Bordes, Y. Bengio, Domain adaptation for large-scale sentiment classification: A deep learning approach, in: Proc. the 28th International Conference on Machine Learning, ICML, Bellevue, WA, 2011, pp. 513–520.
[39] J. Deng, Z. Zhang, F. Eyben, B. Schuller, Autoencoder-based unsupervised domain adaptation for speech emotion recognition, IEEE Signal Process. Lett. 21 (9) (2014) 1068–1072.
[40] J. Deng, X. Xu, Z. Zhang, S. Frühholz, B. Schuller, Semi-supervised autoencoders for speech emotion recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 26 (1) (2017) 31–43.
[41] M. Abdelwahab, C. Busso, Domain adversarial for acoustic emotion recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 26 (12) (2018) 2423–2435.
[42] J. Han, Z. Zhang, N. Cummins, B. Schuller, Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives, IEEE Comput. Intell. Mag. 14 (2) (2018) 68–81.
[43] J. Gideon, M. McInnis, E.M. Provost, Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG), IEEE Trans. Affect. Comput. (2019) 14.
[44] S. Zhang, S. Zhang, T. Huang, W. Gao, Q. Tian, Learning affective features with a hybrid deep model for audio–visual emotion recognition, IEEE Trans. Circuits Syst. Video Technol. 28 (10) (2017) 3030–3043.
[45] H. Kaya, F. Gürpınar, A.A. Salah, Video-based emotion recognition in the wild using deep transfer learning and score fusion, Image Vis. Comput. 65 (2017) 66–75.
[46] C. Sun, L. Huang, X. Qiu, Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence, 2019, arXiv preprint arXiv:1903.09588.

[47] J. Han, Z. Zhang, M. Schmitt, M. Pantic, B. Schuller, From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty, in: Proc. ACM International Conference on Multimedia, MM, Mountain View, CA, 2017, pp. 890–897.

[48] R. Xia, Y. Liu, A multi-task learning framework for emotion recognition using 2D continuous space, IEEE Trans. Affect. Comput. 8 (1) (2017) 3–14.

[49] S. Parthasarathy, C. Busso, Jointly predicting arousal, valence and dominance with multi-task learning, in: Proc. Annual Conference of the International Speech Communication Association, INTERSPEECH, Stockholm, Sweden, 2017, pp. 1103–1107.

[50] Y. Zhang, Y. Liu, F. Weninger, B. Schuller, Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, New Orleans, LA, 2017, pp. 4990–4994.

[51] Z. Zhang, B. Wu, B. Schuller, Attention-augmented end-to-end multi-task learning for emotion prediction from speech, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Brighton, UK, 2019, pp. 6705–6709.

[52] P. Ekman, Basic emotions, in: Handbook of Cognition and Emotion, John Wiley & Sons, Ltd., 1999, pp. 45–60, chapter 3.

[53] J.A. Russell, A circumplex model of affect, J. Personal. Soc. Psychol. 39 (6) (1980) 1161.

[54] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognit. Lett. 125 (2019) 264–270.

[55] P. Tzirakis, G. Trigeorgis, M.A. Nicolaou, B. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, IEEE J. Sel. Top. Sign. Proces. 11 (8) (2017) 1301–1309.

[56] E. Cambria, Affective computing and sentiment analysis, IEEE Intell. Syst. 31 (2) (2016) 102–107.

[57] S.L. Lo, E. Cambria, R. Chiong, D. Cornforth, Multilingual sentiment analysis: From formal to informal and scarce resource languages, Artif. Intell. Rev. 48 (4) (2017) 499–527.

[58] M. McCloskey, N.J. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem, in: Psychology of Learning and Motivation, volume 24, 1989, pp. 109–165.

[59] B. Thompson, J. Gwinnup, H. Khayrallah, K. Duh, P. Koehn, Overcoming catastrophic forgetting during domain adaptation of neural machine translation, in: Proc. the North American Chapter of the Association for Computational Linguistics, NAACL, Minneapolis, Minnesota, 2019, pp. 2062–2068.

[60] A. Chaudhry, P.K. Dokania, T. Ajanthan, P.H. Torr, Riemannian walk for incremental learning: Understanding forgetting and intransigence, in: Proc. the European Conference on Computer Vision, ECCV, Munich, Germany, 2018, pp. 532–547.

[61] J. Schwarz, W. Czarnecki, J. Luketina, A. Grabska-Barwinska, Y.W. Teh, R. Pascanu, R. Hadsell, Progress & compress: A scalable framework for continual learning, in: Proc. International Conference on Machine Learning, ICML, Stockholm, Sweden, 2018, pp. 4535–4544.

[62] X. Liu, M. Masana, L. Herranz, J. Van de Weijer, A.M. Lopez, A.D. Bagdanov, Rotate your networks: Better weight consolidation and less catastrophic forgetting, in: Proc. 24th International Conference on Pattern Recognition, ICPR, Beijing, China, 2018, pp. 2262–2268.

[63] C. Clopath, Synaptic consolidation: An approach to long-term learning, Cogn. Neurodynamics 6 (3) (2012) 251–257.

[64] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, M. Pantic, AVEC 2011 – The first international audio/visual emotion challenge, in: Proc. 1st Annual Workshop on Audio/Visual Emotion Challenge, AVEC, Memphis, TN, 2011, pp. 415–424.

[65] F. Ringeval, A. Sonderegger, J.S. Sauer, D. Lalanne, Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions, in: Proc. 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG, Shanghai, China, 2013, pp. 1–8.

[66] M. Schmitt, B. Schuller, openXBOW—Introducing the Passau open-source crossmodal bag-of-words toolkit, J. Mach. Learn. Res. 18 (96) (2017) 1–5.

[67] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, K. Truong, The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing, IEEE Trans. Affect. Comput. 7 (2) (2016) 190–202.

[68] F. Eyben, M. Wöllmer, B. Schuller, openSMILE – The Munich versatile and fast open-source audio feature extractor, in: Proc. ACM International Conference on Multimedia, MM, Florence, Italy, 2010, pp. 1459–1462.

[69] T. Baltrušaitis, P. Robinson, L.-P. Morency, OpenFace: An open source facial behavior analysis toolkit, in: Proc. IEEE Winter Conference on Applications of Computer Vision, WACV, Lake Placid, NY, 2016, pp. 1–10.

[70] R. Jozefowicz, W. Zaremba, I. Sutskever, An empirical exploration of recurrent network architectures, in: Proc. International Conference on Machine Learning, ICML, Lille, France, 2015, pp. 2342–2350.

[71] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmi, M. Pantic, AVEC 2017–Real-life depression, and affect recognition workshop and challenge, in: Proc. 7th International Workshop on Audio/Visual Emotion Challenge, AVEC, Mountain View, CA, 2017, pp. 3–10.

[72] J. Han, Z. Zhang, N. Cummins, F. Ringeval, B. Schuller, Strength modelling for real-world automatic continuous affect recognition from audiovisual signals, Image Vis. Comput. 65 (2017) 76–86.

[73] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A.A. Rusu, A. Pritzel, D. Wierstra, Pathnet: Evolution channels gradient descent in super neural networks, 2017, arXiv preprint arXiv:1701.08734.

[74] A. Gepperth, C. Karaoguz, A bio-inspired incremental learning architecture for applied perceptual problems, Cogn. Comput. 8 (5) (2016) 924–934.

[75] A.A. Rusu, N.C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, R. Hadsell, Progressive neural networks, 2016, arXiv preprint arXiv:1606.04671.

[76] J. Yoon, E. Yang, J. Lee, S.J. Hwang, Lifelong learning with dynamically expandable networks, in: Proc. International Conference on Learning Representations, ICLR, New Orleans, LA, 2018, p. 11.

[77] Z. Ahmad, R. Jindal, A. Ekbal, P. Bhattachharyya, Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding, Expert Syst. Appl. 139 (2020) 112851.

[78] M.W. Woo, J. Lee, K. Park, A reliable IoT system for personal healthcare devices, Future Gener. Comput. Syst. 78 (2018) 626–640.

[79] I. Azimi, T. Pahikkala, A.M. Rahmani, H. Niela-Vilén, A. Axelin, P. Liljeberg, Missing data resilient decision-making for healthcare IoT through personalization: A case study on maternal health, Future Gener. Comput. Syst. 96 (2019) 297–308.

[80] D. Casado-Mansilla, J. López-de Armentia, D. Ventura, P. Garaizar, D. López-de Ipina, Embedding intelligent eco-aware systems within everyday things to increase people's energy awareness, Soft Comput. 20 (5) (2016) 1695–1711.

[81] D. Casado-Mansilla, I. Moschos, O. Kamara-Esteban, A.C. Tsolakis, C.E. Borges, S. Krinidis, A. Irizar-Arrieta, K. Konstantinos, A. Pijoan, D. Tzovaras, D. López-de Ipina, A human-centric & context-aware IoT framework for enhancing energy efficiency in buildings of public use, IEEE Access 6 (2018) 31444–31456.

[82] Z. Khan, Z. Pervez, A.G. Abbasi, Towards a secure service provisioning framework in a smart city environment, Future Gener. Comput. Syst. 77 (2017) 112–135.

[83] K. Soomro, M.N.M. Bhutta, Z. Khan, M.A. Tahir, Smart city big data analytics: An advanced review, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. (2019) e1319.

Jing Han received her bachelor degree (2011) in electronic and information engineering from Harbin Engineering University (HEU), China, and her master degree (2014) from Nanyang Technological University, Singapore. She is a Ph.D. candidate with the Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg, Germany, involved in two EU's Horizon 2020 projects SEWA and RADAR CNS. Her research interests are related to deep learning for human-centric multimodal affective computing and health care. Besides, she co-chaired the 7th Audio/Visual Emotion Challenge (AVEC) and workshop in 2017, and served as a program committee member of the 8th AVEC challenge and workshop in 2018. Moreover, she was awarded student travel grants from IEEE SPS and ISCA to attend ICASSP and INTERSPEECH in 2018. She (co-)authored more than 30 publications in peer-reviewed journals and conference proceedings.

Zixing Zhang received his master degree in physical electronics from the Beijing University of Posts and Telecommunications (BUPT), China, in 2010, and his Ph.D. degree in computer engineering from Technical University of Munich (TUM), Germany, in 2015. From 2017 to 2019, he was a research associate with the Department of Computing at the Imperial College London (ICL), UK. Before that, he was a postdoctoral researcher at the University of Passau, Germany. To date, he has authored more than 80 publications in peer-reviewed books, journals, and conference proceedings. His research mainly focuses on deep learning technologies for speaker-centred state and health computing. He has organised special sessions, such as at the IEEE 7th Affective Computing and Intelligent Interaction (ACII) conference in 2017 and at the 43nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in 2018. Moreover, he serves as a reviewer

for numerous leading-in-their fields journals and conferences, a programme committee member and an area chair for many international conferences, and a leading guest editor of the IEEE Transactions on Emerging Topics in Computational Intelligence.

Maja Pantic is a Professor of Affective and Behavioural Computing and leader of the i·BUG group and the Research Director of the Samsung AI Centre in Cambridge (SAIC), working on machine analysis of human non-verbal behaviour and its applications to human–computer, human–robot, and computer-mediated human–human interaction. Prof. Pantic published more than 400 technical papers in the areas of machine analysis of facial expressions, machine analysis of human body gestures, audiovisual analysis of emotions and social signals, and human–centred machine interfaces. She has more than 32 000 citations to her work, and has served as an (co-)chair and organisation/program committee member at numerous conferences in her areas of expertise. She is a Fellow of the IEEE.

Björn Schuller received his diploma in 1999, his doctoral degree for his study on Automatic Speech and Emotion Recognition in 2006, and his habilitation and Adjunct Teaching Professorship in the subject area of Signal Processing and Machine Intelligence in 2012, all in electrical engineering and information technology from TUM in Munich/Germany. He is Professor of Artificial Intelligence in the Department of Computing at the Imperial College London/UK, where he heads GLAM — the Group on Language, Audio & Music, Full Professor and head of the Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg/Germany, and CSO of audEERING. He was previously full professor and head of the Chair of Complex and Intelligent Systems at the University of Passau/Germany. Professor Schuller is President-emeritus of the Association for the Advancement of Affective Computing (AAAC), former member of the IEEE Speech and Language Processing Technical Committee, Fellow of the IEEE, and Senior Member of the ACM. He (co-)authored 5 books and more than 800 publications in peer-reviewed books, journals, and conference proceedings leading to more than overall 30 000 citations (h-index = 80). He is general chair of ACII 2019, co-Program Chair of Interspeech 2019 and ICMI 2019, repeated Area Chair of ICASSP, and former Editor in Chief of the IEEE Transactions on Affective Computing next to a multitude of further Associate and Guest Editor roles and functions in Technical and Organisational Committees.