Adapting the Interplay Between Personalized and Generalized Affect Recognition Based on an Unsupervised Neural Framework

Pablo Barros, Emilia Barakova, and Stefan Wermter

Abstract—Recent emotion recognition models, most of them being based on strongly supervised deep learning solutions, are rather successful in recognizing instantaneous emotion expressions. However, when applied to continuous interactions, these models show a weaker adaptation to a person-specific and long-term emotion appraisal. In this article, we present an unsupervised neural framework that improves emotion recognition by learning how to describe continuous affective behavior of individual persons. Our framework is composed of three self-organizing mechanisms: (1) a recurrent growing layer to cluster general emotion expressions, (2) a set of associative layers, acting as *affective memories* to model specific emotional behavior of individual persons, (3) and an online learning layer which provides contextual modeling of continuous emotion expressions. We propose different learning strategies to integrate all three mechanisms and to improve the performance on arousal and valence recognition of the OMG-Emotion dataset. We evaluate our model with a series of experiments ranging from ablation studies assessing the different contributions of each neural component to an objective comparison with state-of-the-art solutions. The results from the evaluations show a good performance on emotion recognition of continuous emotions on monologue videos. Furthermore, we discuss how the model self-regulates the interplay between generalized and personalized emotion perception and how this influences the model's reliability when recognizing unseen emotion expressions.

Index Terms—Emotion recognition, personalized emotion perception, online learning, unsupervised learning, continual learning

1 Introduction

Although it is widely accepted that basic emotional concepts are perceived by different persons around the world in a consistent manner [1], applying a computational model able to recognize emotions in natural scenarios remains a difficult task [2]. A major challenge identified by Cavallo *et al.* [2] is that each person might express affect differently, sometimes transitioning between several basic emotions in a short time period to express a complex emotional state. In addition, both understanding and display of emotion can vary depending on the situation, the interaction partner and even on the time of the day [3], [4]. In an attempt to create a system able to recognize individual and continuous emotions, several researchers defined more complex emotional states, such as confusion, surprise, and concentration [5], [6], [7], which increases the complexity of the recognition models.

To deal with the newly introduced complexity, most of the recent solutions on emotion recognition employ an extreme generalization approach, usually based on end-to-end deep learning techniques [8]. Such models usually learn how to represent affective features from a large number of data samples, using strongly supervised methods [9], [10], [11], [12]. As a result, these models can extract audio-visual features from a collection of different individuals, and maximize the generalization on emotion expression recognition. The development of such models was supported by the collection of several "in-the-wild" datasets [13], [14], [15], [16] which provided large amounts of strongly labelled data.

These datasets usually contain emotion expressions from various web sources in the form of short instances ranging from single frames to a few seconds of video material. Because of the large availability of data to train with, the performance of deep learning-based solutions provides the state of the art achievements when benchmarked on these datasets [17], [18], [19], [20]. When applied to real-world problems, however, such deep learning models tend to perform poorly [21], [22]. In addition, most deep learning solutions although capable of representing audio and visual affect from a wide range of persons, can only categorize simple representations of emotional display, such as positive or negative expressions [23], [24].

The most common cause of the poor performance on real-world scenarios is the high dependence of such solutions on learning the separation boundaries of the emotion categories based on the examples of the training set [2], [25]. Emotion categories which have more examples in the training dataset will enforce a strong bias on the model learning. Also, although most of these datasets contain millions of examples, they are still much limited by providing short instances of emotional expressions [26]. These models are successful in recognizing short representations of emotions

- P. Barros is with the Cognitive Architecture for Collaborative Technologies Unit (CONTACT), Italian Institute of Technology (IIT), 16132 Genova, Italy. E-mail: pablo.alvesdebarros@iit.it.
- E. Barakova is with the Department of Industrial Design, Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands. E-mail: e.i.barakova@tue.nl.
- S. Wermter is with Knowledge Technology, Department of Informatics, University of Hamburg, 20146 Hamburg, Germany. E-mail: wermter@informatik.uni-hamburg.de.

Manuscript received 5 November 2019; revised 3 June 2020; accepted 4 June 2020. Date of publication 15 June 2020; date of current version 6 September 2022.
(Corresponding author: Pablo Barros.)
Recommended for acceptance by A. A. Salah.
Digital Object Identifier no. 10.1109/TAFFC.2020.3002657

on the specific set of persons/situations which were present in each of the training datasets [27]. As soon as a new expression is perceived, or a person with characteristics that were not well represented in the training set expresses an emotion, these models tend to have a poor performance.

Recent neural models based on adversarial learning [28], [29] overcame the necessity of strongly labelled data, however, they tend to need even more examples from different persons, thus much more data points are needed. In a recent work, we explored a similar approach for generating personalized emotion expressions [30]. Our model was successful in using the generated data to predict arousal and valence, but only on single images, which limits the models' application in continual interactions. It also demanded an enormous training effort to optimize a large number of parameters in order to stabilize the adversarial learning. Existing solutions address the problem of continual learning in deep learning models by introducing transfer learning techniques [21], [31], neural activation and data distribution regularization [32], [33], and the unsupervised learning of affective features [34], [35]. Most of these models present an improvement of performance when evaluated on specific datasets, but maintain the same limitations when applied to real-world scenarios, as the adaptation process is expensive and slow, demanding many interactions to learn new data instances.

Previously, we proposed a developmental approach for emotion expression recognition [36], [37] which addressed the problem of online adaptation of emotional categories to newly perceived expressions. This model implemented self-organizing layers which create clusters of similar expressions based on audio/visual characteristics extracted from a convolutional neural network. The model learned, online, how to correlate newly perceived expressions with the learned clusters, and how to create new clusters to represent new perceived emotion expressions when needed. It achieved state-of-the-art performance on recognizing instantaneous emotion expressions by identifying to which cluster a newly perceived expression belongs to, without the need to re-train the entire convolutional neural network. One of the limitations of our previous work was the pre-defined number of neurons and the topological arrangement of the self-organizing network. This constrained the number of emotion concepts the model learned and created artificial relations on how the emotion expressions were represented, which influenced its application to unconstrained scenarios. Also, the neurons only learned how to represent short instantaneous expressions, which limited its application to natural scenarios and it did not allow the model to adapt to the perceived expressions while an interaction happens.

In this paper, we address the problem of learning adaptable emotional representations by focusing on improving emotion recognition based on two characteristics: the universal generalization of human emotional expressions and the individualization of affective understanding. A person can express sadness by crying, screaming, or even smiling [38], which makes it hard to be categorized in a computational model. The diversity of emotion expressions becomes even more complex when we take into consideration interpersonal characteristics such as cultural background, personality, and genetics [39]. Deep learning models have been shown to be expert on providing generalization, and for representing these characteristics of unknown persons. However, such models have difficulty to adapt to the individualized characteristics of a subject, as they rely on learning from multiple examples. When humans already know a specific person, they can adapt their own perception to how that specific person expresses emotions [3]. The interplay between recognizing an individual's facial characteristics, i.e., how certain muscles in the face are moved, mouth and eye positions, blushing intensity, and clustering them into emotional states is one of the key points for modeling a human-like emotion recognition system [40].

To alleviate the above-mentioned limitations, we propose the Affective Memory Framework (*AffMem*), explained in detail in Section 2, as an adaptable solution for recognition of continuous emotional expressions. The *AffMem* builds on our previous work by introducing an interplay between general affective perception and the online learning of individualized emotion characteristics. We extract affective features from audio/visual expressions using a strongly pre-trained Cross-Channel Convolutional Neural Network (*CCCNN*) [36], which showed to represent multimodal affect reliably. Once the extracted features have been dealt with, we address the problems of end-to-end learning by providing a modular mechanism which integrates three growing self-organizing mechanisms. The first mechanism is the *General-GWR*, which is inspired by our previous work [37] and implements a recurrent Growing-When-Required network [41] that learns how to represent general emotion expressions using unsupervised clustering of audio/visual characteristics. The second mechanism implements a series of GWR networks, the *affective memories*, to learn, online, how to cluster the audio/visual characteristics of the expressions from a single person in order to learn specific individualized relations between them. The third, and last mechanism, the Mood, integrates both the *General-GWR* and the *affective memories* into a continual learning stateful neural network that represents perceived expressions.

In Section 3, we summarize our efforts to perform an objective evaluation of the proposed framework. We evaluate it with two different sets of experiments: an ablation study, which uses three different unisensory and multisensory emotion expression datasets to assess the contribution of the proposed individual neural mechanism, and an integrative experiment to assess the framework's performance on recognizing continuous emotion expressions. During the experiments, we also take into consideration the information flow of the different neural mechanisms when recognizing longitudinal emotion expressions. The results of these experiments, presented in Section 3.3, show that our model outperforms the current state-of-the-art models on continuous emotion expression recognition. We, thus, present in Section 4 an overview of the different abstraction levels learned by the proposed model, explaining how they contribute to the final performance of the model. Finally, in Section 5, we conclude that our model presents an improvement in continuous emotion expression recognition models by providing an online learning mechanism for modeling the interplay between general and personalized affective behavior.

2 The Affective Memory Framework (AffMem)

The Affective Memory Framework addresses the problem of learning adaptable affective representations by proposing

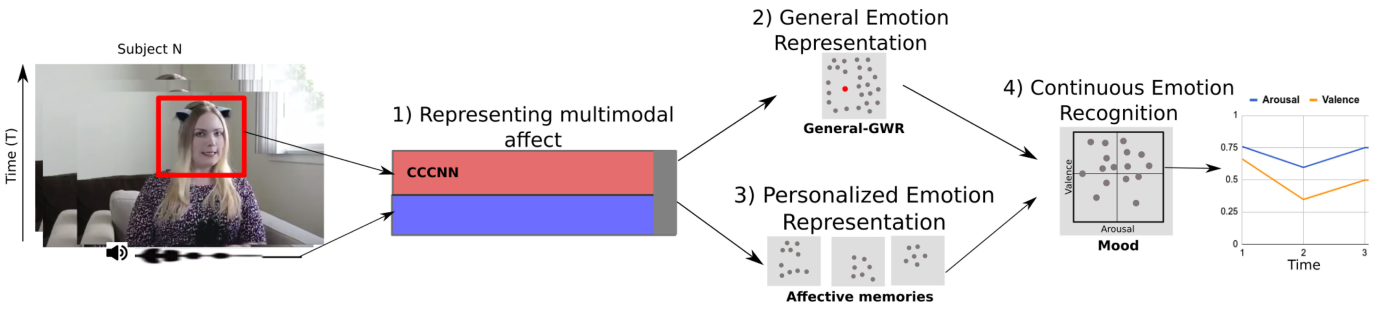
Subject N
Time (T)
1) Representing multimodal affect
CCCNN
2) General Emotion Representation
General-GWR
3) Personalized Emotion Representation
Affective memories
4) Continuous Emotion Recognition
Valence
Arousal
Mood
Arousal
Valence
1
0.75
0.5
0.25
0
1
2
3
Time

Fig. 1. An overview of the Affective Memory Framework (*AffMem*). It integrates 1) the representation of expressions using a Cross-channel Convolutional Neural Network (*CCCNN*), 2) the generalized emotion recognition with the General-GWR, 3) the personalized affective representation using the *affective memories*, and 4) the Mood which integrates generalized and personalized representation into a continuous emotion recognition.

and integrating three mechanisms: first, it clusters the perceived expressions into general emotional concepts with the *General-GWR*; second, it learns online, and in parallel with the *General-GWR*, how to cluster the specific characteristics of the emotion expressions of individual persons into emotional concepts using the *affective memories*; and third, it creates a continuous reading of arousal and valence which integrates both generalized and personalized perception through the *Mood* module. To represents multimodal emotion expressions we use a pre-trained *Cross-channel Convolutional Neural Network*. Fig. 1 illustrates the *AffMem* and its process flow.

Each of the proposed neural modules addresses one individual problem of emotion recognition. Together, they contribute with a learning flow that balances the interplay between generalized and personalized emotion perception which results in a dynamic system that 1) *represents multisensory affect*, 2) *classifies general emotion expressions* into known clusters, 3) learns *individualized emotion representations* and 4) provides *continuous emotion recognition*.

2.1 Representing Affective Stimuli

The first necessary step to automatically process affective stimuli is to find an appropriate represention. Our framework provides such representation by pre-training the Face Channel and the Auditory Channel of the Cross-channel Convolutional Neural Network [36]. The *CCCNN*, illustrated in Fig. 2, implements single-modality convolutional neural channels to learn affective descriptors which showed to be competitive with state-of-the-art models in the recognition of instantaneous emotional expressions while maintaining a very light-weighted architecture.

We updated the original Face channel to follow a topology based on the VGG16 model [42], but with much fewer parameters, which made it much more robust towards facial-expression representation. The Face channel has 10 convolutional layers, including 4 pooling layers. We use batch normalization within each convolutional layer and a dropout function after each pooling layer. Following the original *CCCNN* architecture, we apply shunting inhibitory fields [43] in our last layer. Each shunting neuron S_{nc}^{xy} at the position (x, y) of the nth receptive field in the cth layer is activated as

$$S_{nc}^{xy} = \frac{u_{nc}^{xy}}{a_{nc} + I_{nc}^{xy}}, \tag{1}$$

where u_{nc}^{xy} is the activation of the common unit in the same position and I_{nc}^{xy} is the activation of the inhibitory neuron. A learned passive decay term, a_{nc} is the same for each shunting inhibitory field. Each convolutional and inhibitory layer of the Face Channel implements a ReLu activation function.

The Auditory channel learns how to represent prosodic information from speech to recognize emotional expressions. In this regard, it receives as a Mel-Frequency Cepstral Coefficients (MFCCs) representation. When transformed to MFCCs, the auditory signals lose their spacial locality which makes the common 2D convolutions, not suitable processing approach. Abdel-Hamid *et al.* [44] propose the use of 1D filters to solve this problem. We adopted this approach and implemented the Auditory channel as a series of 1D convolutions which are applied to the Y-axis of the MFCC spectrum. The

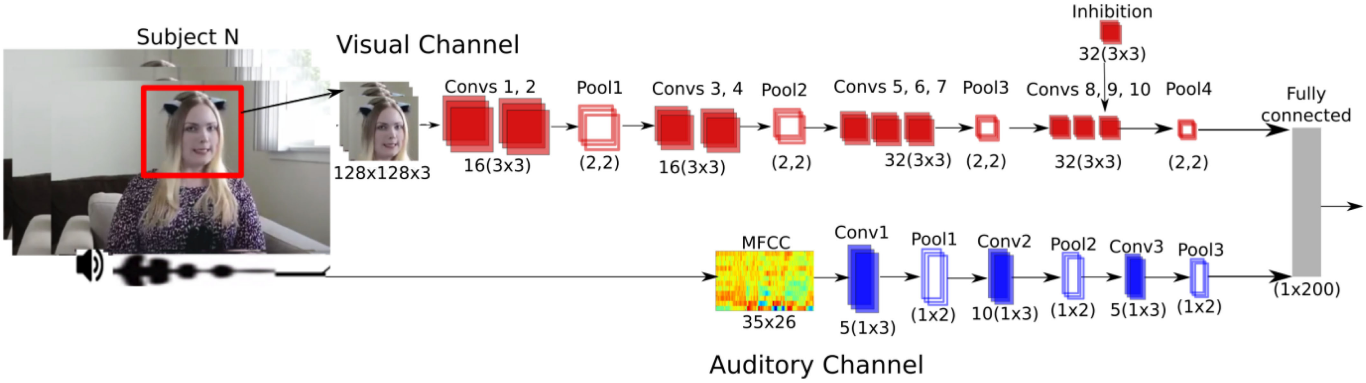

Fig. 2. Detailed architecture and parameters of the Visual and Auditory channels of our *CCCNN* architecture.

filters will learn how to correlate the representation per axis and not within neighbors, which does not carry any correlation. Pooling is also applied in one dimension after each convolutional layer, always keeping the same topological structure. The auditory channel is composed of three layers, each one with one-dimensional filters and followed by a pooling layer.

As typical for most deep learning models, our *CCCNN* has several hyperparameters to be tuned. We optimized our model to maximize the recognition accuracy using a Tree-structured Parzen Estimator (TPE) [45] and use the optimal training parameters through all of our experiments. The results of our optimization indicated that training the model in two steps yielded the best recognition performance. First, we train the individual channels using specific unimodality corpora, and then we fine-tune the model by concatenating the output of each channel and re-training it with an audio/visual emotion expression dataset. In both cases, the output of the model is fed to a fully connected layer with 200 units, each one implementing a ReLu activation function, which is then fed to an output layer. Our model is trained using a categorical cross-entropy loss function.

2.2 General Emotion Expressions

The *CCCNN*, like many other purely deep learning solutions, are enforced to learn a hierarchical representation of emotion expressions. After a strong supervised training procedure, these networks create separation boundaries to classify the perceived affect into previously known emotional concepts. As soon as a new expression or even a known expression expressed by a different person is perceived, the *CCCNN* relies on an extreme generalization to recognize it. If the expression is not represented in the training set, the performance is usually very low. Re-training the network to improve its recognition towards specific expressions is a very expensive process and requires a large number of new training samples.

In our previous work [36], we addressed the problem of adapting a convolutional neural network towards novel data by endowing it with an unsupervised layer which learns the separation boundaries based on the clusters of a self-organizing map (SOM). The advantage of this approach was to use the expertise of a deep neural network to describe expressions, as a pre-trained feature extractor, together with the unsupervised learning of clusters representing expressions, without the necessity of re-training the entire network every time a new expression, or person, was perceived. The self-organizing network adapted its existing neurons, using the Hebbian learning-based neighborhood function, to accommodate a novel expression. The hybrid model achieved better performance when compared to purely supervised convolutional neural networks when recognizing general expressions in cross-database experiments. The model presented, however, serious scalability problems, as it had a pre-defined number of neurons and topological structure which did not let it learn in unconstrained scenarios.

To address the limitations of our previous work, we propose in this paper the *General-GWR* layer which creates prototype neurons from the multimodal emotion expressions represented by the *CCCNN* based on Growing-When-Required Networks (GWR). The GWR is a self-organizing network which learns how to represent data by training prototype neurons through an unsupervised Hebbian learn update rule. Recent studies [46], [47] investigate the successful employment of such networks into continual learning for different high-abstraction tasks. They discuss how the GWR adapts to the input data much faster, and using fewer resources than traditional lifelong and transfer learning mechanisms [48]. Each neuron on the GWR grid represents an approximation of a series of perceived data points so that a newly perceived data point can be represented by one or more neurons on the network. By clustering these neurons on known concepts, such as arousal and valence, we can use them to characterize the newly perceived expressions.

To improve the capability of dealing with longer emotion expressions, we implement gamma connections [41] to our *General-GWR*. The gamma connections are inspired by the gamma memory models [49] and embed the prototype neurons with a context layer. The context layer will encode the representation of a short-sequence of expressions. Our assumption is that the gamma connections would improve the robustness of the prototype neurons by reducing the sensitivity of the neuron to learn outlier representations.

The *General-GWR* associates a best-matching unit (BMU) b with an input taking into consideration the activity of the network for the current input and previous inputs, the latter represented as a temporal context via the gamma connections. Each neuron of the *General-GWR* consists of a weight vector w_j and a number K of global context descriptors $c_{j,k}$ (with $w_j, c_{j,k} \in \mathbb{R}^n$). As a result, each neuron on the network will encode prototype sequence-selective representations of the input taking into consideration each neuron's representation but also a each neurons' context. Given a set of N neurons, b is computed with respect to the input $x(t) \in \mathbb{R}^n$ as

$$b = \arg\min_{j \in N}\big(d_j\big),$$
$$d_j = \alpha_0 \|x(t) - w_j\|^2 + \sum_{k=1}^{K} \alpha_k \|C_{j,k}(t) - c_{j,k}\|^2, \tag{2}$$

$$\mathbf{C}_{j,k}(t) = \beta \cdot w_{j(t-1)} + (1 - \beta) \cdot C_{j,k-1}(t-1), \tag{3}$$

where d_j represents the distances between the input stimuli and all the neurons of the *General-GWR*, α_i and $\beta \in (0;1)$ are constant values that modulate the influence of the current input with respect to previous neural activity, $w_j(t-1)$ is the weight of the winner neuron at $t-1$, and $C_{j,k}(t) \in \mathbb{R}^n$ is the global context of the network ($C_{j,k}(t_0) = 0$).

New connections are created between the BMU and the second-BMU with relation to the input. When a BMU is computed, all the neurons it is connected to are referred to as its topological neighbors. Each neuron is equipped with a habituation counter $h_j \in [0,1]$ expressing how frequently it has been fired based on a simplified model of how the efficacy of a habituating synapse reduces over time.

The habituation rule is given by $\Delta h_j = \tau_j \cdot \kappa \cdot (1 - h_j) - \tau_j$, where κ and τ_j are constants that control the decreasing behavior of the habituation counter [50]. The habituation is needed in our scenario in order to enforce the network to update its prototype neurons with a higher impact towards newly perceived affective input. To establish whether a

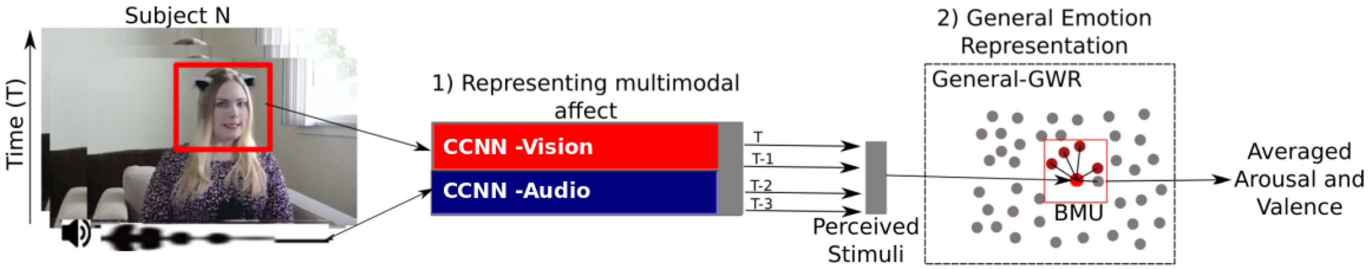
Subject N
Time (T)
1) Representing multimodal affect
CCNN -Vision
CCNN -Audio
T
T-1
T-2
T-3
Perceived Stimuli
2) General Emotion Representation
General-GWR
BMU
Averaged Arousal and Valence

Fig. 3. The pre-trained *General-GWR* processes a perceived input, composed always of the current perceived stimulus and three previous stimuli. The Best-Matching Unit (BMU) is computed together with its five closest neighbors, and the final averaged arousal and valence is calculated.

neuron is habituated, its habituation counter h_j must be smaller than a given habituation threshold t_h.

The network is initialized with two neurons and, at each learning iteration, it inserts a new neuron whenever the activation of the network for the expression i, $a(i) = \exp(-d_j)$, is smaller than a given threshold t_a, i.e., a new neuron is created if $a(t) < t_a$ and $h_j < t_h$. Each neuron also contains an aging mechanism so as soon as a neuron is activated as a BMU or a BMU neighbor, its aging counter is set to zero. If not, its aging counter is increased. If a neuron has an aging counter higher than a threshold t_o, it will be removed. A neuron which is not activated often will be removed, and this mechanism is extremely important to avoid outliers on the creation of prototypes by the network.

The training of the neurons is carried out by adapting the BMU b and its topological neurons n according to

$$\Delta \mathbf{b} = \epsilon_i \cdot h_i \cdot (\mathbf{x}(t) - \mathbf{b}), \tag{4}$$

$$\Delta \mathbf{c}_{i,k} = \epsilon_i \cdot h_i \cdot (\mathbf{C}_k(t) - \mathbf{c}_{i,k}), \tag{5}$$

where ϵ_i is a constant learning rate.

To allow the *General-GWR* to perform classification of emotion expressions, we implement associative labeling [41] to each neuron. During the training phase, we assign to each neuron two continuous values, representing arousal and valence. If a training sample has arousal (l_a) and valence (l_v) associated with it, we update the labels of the BMU ($l_{i,a}$ for the associated arousal and $l_{i,v}$ for the associated valence) with

$$l_{i,a} = l_{i,a} + (l_a - l_{i,a}) \times \gamma_l \tag{6}$$

$$l_{i,v} = l_{i,v} + (l_v - l_{i,v}) \times \gamma_l, \tag{7}$$

where γ_l is a labeling learning factor and it is defined during training.

To categorize a newly perceived expression, we just read the labels of the BMU associated with it. Important to note that although we use the association labeling in our training, the training of the neurons is purely unsupervised. The *General-GWR* will keep learning even if there is no arousal or valence associated with a training example.

The learning process of the *General-GWR* is unsupervised and driven by the examples of the training data. The *General-GWR* either allocates new neurons or adapts existing ones in response to novel input. To increase the robustness of the *General-GWR* against noisy information, we use a weighted average of the labels of the BMU and up to five of its closest neighbors to calculate the final arousal and valence. Fig. 3 illustrates the *General-GWR* recognition process flow. The *General-GWR* is also pre-trained with "in-the-wild" datasets, so it learns prototype neurons with general affective concepts.

The hyperparameters of the network listed in Table 1, were optimized to reflect an optimal performance on cross-dataset emotion recognition. To optimize the parameters of the *General-GWR* we also used a Tree-structured Parzen Estimator [45]. The *General-GWR* contributes to our previous work by providing the growing and shrinking mechanisms which allows it to learn novel prototype neurons without any topological restriction. Also, the network creates neighboring connections between similar prototype neurons, which means that neighbor neurons will represent similar expressions (i.e., an emotional concept). As the neurons are approximations of perceived expressions, neuron neighborhoods will represent similar emotional concepts. In contrast to a common deep learning network, the amount of data needed to create new neighborhoods, and in this case, the separation boundaries of emotional concepts, is much smaller. So an emotion expression that is under-represented in the training data will still be well-represented by a prototype neuron. This process also allows the *General-GWR* to distinguish better between noisy information and under-represented expressions as noisy information will generate neurons which are not activated as often as perceived expressions, and soon will be removed. This process was observed for recent continual learning and lifelong learning scenarios involving GWRs [51].

2.3 Individualized Emotion Representations

Our model tackles emotion recognition based on two affective perception perspectives: the pre-disposition of perceiving general emotional concepts, and the online adaptation towards individualized characteristics of affect. A pre-trained *General-GWR* recognizes general affective stimuli by identifying Best-Matching Units (BMUs) which represent the perceived stimuli.

TABLE 1
Training Parameters of the *General-GWR* Optimized to Maximize the Performance of Emotion Expression Recognition

Parameter	Value
Epochs	20
Activity threshold (t_a)	0.1
Habituation threshold (t_h)	0.03
Context size (k)	3
Gamma modulation (α_i and β)	0.64391426, 0.23688282
Habituation modulation (τ_i and κ)	0.08714432, 0.0320586
Labeling factor (γ_l)	0.3

TABLE 2
Training Parameters of Each of the Networks That Belong to the *Affective Memory* Module

Parameter	Value
Epochs	10
Activity threshold (t_a)	0.5
Habituation threshold (t_h)	0.3
Context size	3
Gamma modulation (α_i and β)	0.64391426, 0.23688282
Habituation modulation (τ_i and κ)	0.08714432, 0.0320586
Labeling factor (γ)	0.4

In order to address the individualization problem, we propose the *affective memories*. In contrast to the *General-GWR*, the *affective memories* are GWRs which are trained online to create clusters of perceived expressions that stem from one single person.

GWRs have shown convincing results in encoding changing concepts in online learning tasks [52], which is much related to personalized emotion perception as it can be seen as a specific composition of a general affective display. For instance, for certain individuals, at first, we can recognize that a person is smiling, and associate it with a happy expression. However, after interacting with this person for a longer period of time, we can identify that this specific smile only appears when this person is feeling nervous. Such a process can be properly encoded by GWR networks.

By online training, the *affective memories* can learn specific characteristics of the transitions of the displayed emotions. The first neurons the network creates will be related to the first impression the network has of that person. As the gamma context is updated over time, the new neurons will be impacted, for example, by a transition from a neutral emotion to a happy one. Each new neuron that the network creates represents a different transition of how that person expresses emotions in a much more specific way than the general prototype neurons of the *General-GWR*. For optimal use, there will be one *affective memory* for each person the framework is used on.

To allow the *affective memories* to learn fast and reliable representations, we use a higher activity threshold when compared to the *General-GWR*, as illustrated by the *affective memories* parameters in Table 2. This will, however, make the first expressions dominate the creation of new neurons. Following expressions, however, will only have an impact on the neural update if they happen for a longer period of time. Expressions with extreme arousal and/or valence and that appear only in a short duration would be considered as noise and would not impact the neuron updates. To address this problem, we propose a novel modulation on the update of the neurons based on the neural activation:

$$\Delta \mathbf{w}_j = \epsilon_i \cdot h_i \cdot (\mathbf{x}(t) - \mathbf{w}_j) \cdot (1 - a(i)), \tag{8}$$

$$\Delta \mathbf{c}_{i,k} = \epsilon_i \cdot h_i \cdot (\mathbf{C}_k(t) - \mathbf{c}_{i,k}) \cdot (1 - a(i)). \tag{9}$$

If the neural activation is low, meaning the perceived emotion expression is different from the ones represented by the prototype neurons, the newly perceived expression will have a higher impact on the update of existing neurons. As the *affective memories* learn online, every single perceived expression counts, different from the *General-GWR* which learns based on the extreme generalization of the expressions. The expressions used to train the *affective memories* do not have a label, so we associate the input stimuli with labels coming from the *General-GWR*. This way we enforce the prototype neurons learned by the *affective memories* to carry on the reliability on labeling expressions from *General-GWR* even when the interaction is at the beginning.

The online processing of the *affective memories* happens first with the unsupervised training based on the perceived stimuli for 10 epochs. After training, we calculate the BMUs and five of their neighbors, similar to the *General-GWR*, and provide an averaged arousal and valence. As the learning happens fully unsupervised, and labels are updated online using the ones obtained from the *General-GWR*, the *affective memories* do learn on a closed-loop scenario, without the need of an external teaching signal. Fig. 4 illustrates the training and recognition process of the *affective memories*.

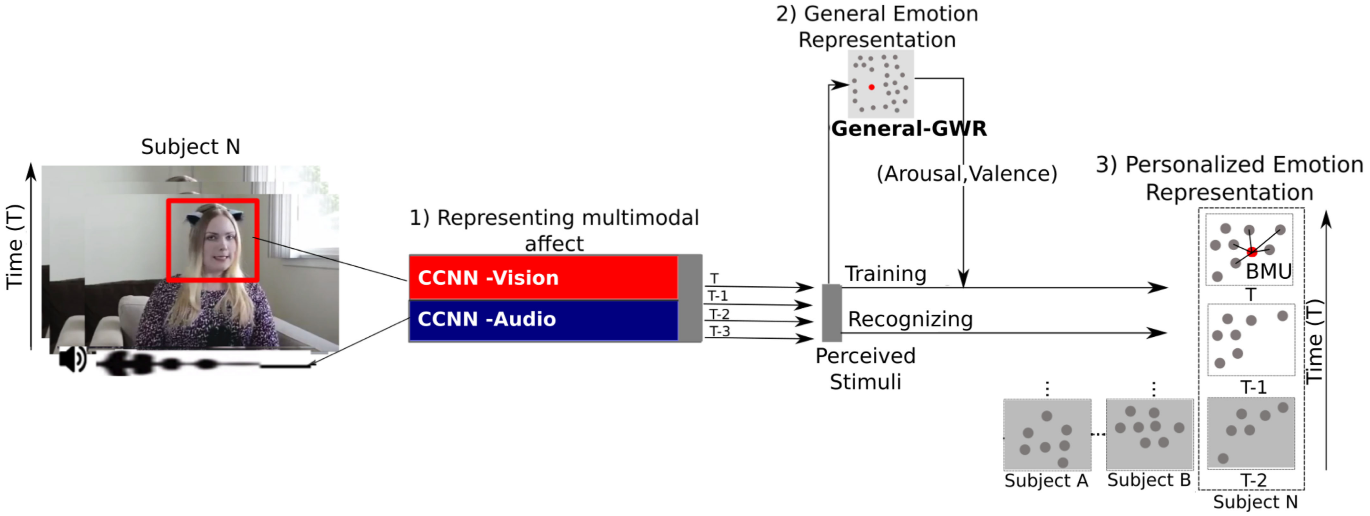

Fig. 4. Training and recognition of the *affective memories*. One *affective memory* is created to learn the perceived stimuli from a single subject using our novel update rule. The online learning uses the labels from the *General-GWR* in the association learning process. After training for 10 epochs, the *affective memories* calculates the arousal and valence of the Best-Matching Unit and five of its neighbors.

The continuous training of the *affective memories* has two goals: 1) to create, online, specific prototype neurons that represent the expressions of one single person, and 2) to be able to improve over time the description of affective behavior of that person. In contrast to the *General-GWR*, the *affective memories* are always learning and always modifying its learned prototypes. With the new neuron update rule, even an expression which was just perceived a few times will be represented. Our assumption is that, once the *affective memories* learned with enough examples of that person, the prototype neurons will be much more robust to represent the particularities of that person, when compared to the ones on the *General-GWR*.

2.4 Continuous Emotion Perception

Both *General-GWR* and the *affective memories* modules run in parallel. Both can describe instantaneous affective behavior through the labels associated with the calculated BMUs. That means we have at our disposition two sets of arousal and valence predictions, one coming from the *General-GWR* and one coming from the *affective memories*. Each of these sets represents an opinion, a generalized and personalized, on how that person is expressing emotions. In order to obtain one reliable prediction, we integrate these values by providing a continuous affective reading. The integration is realized by the third and last mechanism of the *AffMem*, the *Mood*.

The *Mood* is implemented as a GWR network, but different from the other two mechanisms, it does not describe affect by the use of a calculated BMU. The *Mood* is trained as a state-like network, where the representation of perceived affective expressions comes from reading the state of all its neurons at once.

The *Mood* is trained online and continuously, so the *Mood* will first create neurons which will encode the expressions of the first seconds of the interactions. Once the interaction continues, new neurons will be created, or current neurons will be updated, to represent the perceived expressions, and the old neurons will be less updated until they are erased. Each neuron on the *Mood* encodes a direct arousal and valence value instead of being a sensory descriptor. By averaging the weights of all the neurons at any given time, the Mood represents the affective content of the interaction until that point. To avoid the problem of older neurons having a high impact on the arousal and valence reading, we use a weighted average based on the neuron's age. Old neurons, which represent expressions that were perceived at the beginning of the interaction, will impact the affective reading of the Mood less than new neurons.

The recurrent gamma memory provides a context mechanism to train the prototype neurons of both the *General-GWR* and the *affective memories* with short temporal sequences. Both these mechanisms enforce these networks to cluster short-time transitions of emotion expressions instead of one instantaneous snapshot of expression at a time, which would happen in a GWR without the gamma memory. The Mood, however, does not need gamma connections. We can model the dynamics of continuous perception by simply reading all the neurons at different time steps. As soon as there is an affective transition on the perceived interaction, the Mood will create or remove neurons accordingly, and the aging average gives us a simple mechanism to read these dynamics at any given time.

The prototype neurons of the *General-GWR* store general characteristics of the emotional expressions. They can be understood as a first impression, shaped by social experience, of the affective meaning of a perceived expression. The neurons of the *affective memories* encode more specific characteristics, which are related to how one single person expresses emotions. However, the neurons of the *affective memories* are only reliable after they learned from a number of different examples. To integrate the reliability of the calculated BMUs of the *General-GWR* and the dynamics of the *affective memories* on learning specific prototype neurons for that one person, we introduce here a novel update rule for the Mood neurons

$$\Delta \mathbf{w}_i = \epsilon_i \cdot h_i \cdot (\mathbf{av}(t) - \mathbf{w}_i) \cdot (1 - a(t)), \tag{10}$$

where in the first step $a(t)$ is the activation of the *General-GWR* and $av(t)$ is the associated arousal valence, and in the second step $a(t)$ is the activation of the *affective memories* and $av(t)$ the associated arousal and valence. The network activation gives us an objective measure of how well each of these mechanisms can represent an expression. Our hypothesis is that when the *affective memories* are still learning, the *General-GWR* will dominate the Mood updates. However, as soon as the *affective memories* start to learn robust prototypes, their activation will decrease, and as a result they will have a higher impact on the Mood update. Fig. 5 illustrates the process flow of the *Mood*.

3 Evaluating AffMem

To evaluate and understand better the performance of the *AffMem* and the individual impact of each neural mechanism, we perform two different types of experiments. First, we run a series of ablation studies to asses the *quality of the emotion expression representation* (Exp 1.1), the *robustness of the prototype neurons of the General-GWR* (Exp 1.2) and *the role of the dynamic learning of the affective memory* (Exp 1.3). Second, we run a *continuous emotion recognition* experiment (Exp 2) with the entire framework to evaluate its performance on continuous emotion recognition.

3.1 Datasets and Pre-Processing

Our framework relies heavily on a robust emotion expression representation. To train the *CCCNN*, and assess its performance on instantaneous emotion expression recognition, we used two datasets. The AffectNet [14] was used to train and evaluate the Visual channel and it is composed of more than 400 thousand "in-the-wild" images which were manually annotated with seven categorical labels and continuous arousal and valence. The AffectNet contributes to the robustness of the Visual training by providing balanced arousal and valence distribution, as exhibited in Fig. 6 A. As the test set labels are not available to the public, all our experiments were performed using the training and validation samples.

To train the Auditory channel, we used the RADVESS dataset [53]. It contains 7,365 videos of 24 professional actors speaking lexically-matched sentences while performing 7 different emotions (calm, happy, sad, angry, fearful, surprise, and disgust). Each video was annotated on a phrase level,

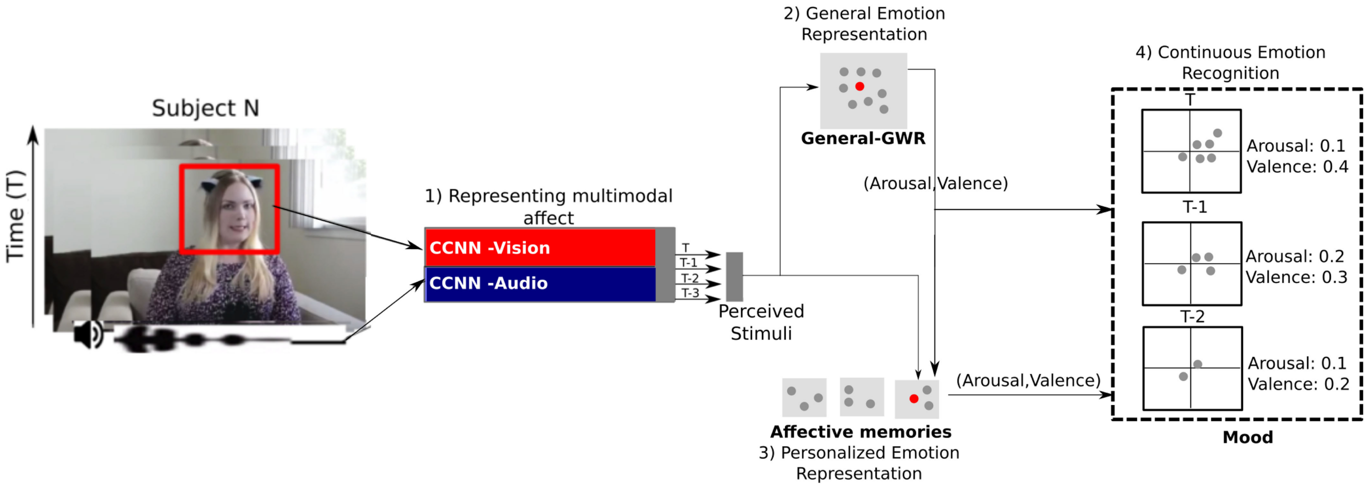

Fig. 5. Continual learning of the state-like Mood. It trains with the arousal and valence values from both *General-GWR* and *affective memories*. At each time-step, new neurons are added or removed from the Mood to represent the perceived emotions. By reading all the Mood neurons, we obtain a snapshot arousal and valence on that time-step.

which provides an even data distribution, as exhibited in Fig. 6 B.

Finally, our multimodal and continuous experiments are performed using the One-Minute Gradual-Emotional Behavior dataset (OMG-Emotion) [26] which contains around 600 Youtube videos which are about a minute in length and are annotated taking into consideration a continuous emotional behavior. The videos were selected using a crawler technique that uses specific keywords based on long-term emotional scenes such as “monologues”, “auditions”, “dialogues” and “emotional scenes”, to guarantee that there is a gradual change in the way the emotions are expressed. As each video contains only one person performing the scene, personalization plays an important role when classifying the person’s emotion expression. The videos were annotated using dimensional arousal and valence, and cover most of the extreme range of arousal and valence values as demonstrated in Fig. 6 C. The dataset has a pre-defined separation between training and testing subsets, and the same person was not present on the training or testing video sets.

For all the evaluated datasets, we follow the training/test separation protocols as described by the dataset authors in order to maintain the comparability with other proposed models. We pre-process the individual stimuli in order to feed them to the *CCCNN*. For each video frame, we detect the face of the subject using the Dlib library [54]. Each face is then resized to a dimension of 128x128. We use audio clips with 1s as input, and each clip is re-sampled to 16000 Hz. We compute the MFCCs of the audio clip and feed it to the Auditory channel of the *CCCNN*. The MFCCs are computed over a window of 25 ms with a slide of 10 ms. We use a frequency resolution of 1,024, which generated a representation with 35 bins, each one with 26 descriptors.

3.2 Metrics

To measure the performance of our experiments, we use two metrics: accuracy, to recognize categorical emotion expressions, and the Concordance Correlation Coefficient (CCC) [55] between the outputs of the models and the true label to recognize arousal and valence. The CCC is computed as

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \quad (11)$$

where ρ is the Pearson’s Correlation Coefficient between model prediction labels and the annotations, μ_x and μ_y denote the mean for model predictions and the annotations and σ_x^2 and σ_y^2 are the corresponding variances. The CCC metric allows us to have a direct comparison with the annotations available on the OMG-Emotion dataset. The use of CCC as the main objective measurement allow us to take into consideration the subjectivity of the perceived emotions

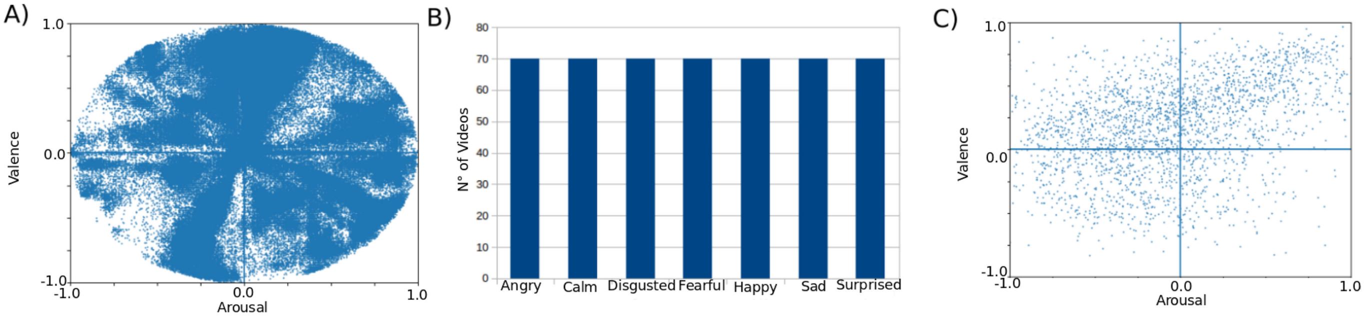

Fig. 6. Illustration of the annotation distribution of the three datasets used in this article: A) the AffectNet dataset [14] has a high variance on arousal and valence and a large number of data points, B) the audio samples of the RAVDESS dataset [53] have a well-balanced data distribution for all emotional classes, and C) the continuous expressions of the OMG-Emotion videos cover a high arousal and valence spread.

for each individual annotator when evaluating the performance of our models.

The Affectnet dataset has 1 label associated per each frame, so we calculate the accuracy directly at the frame level. The RAVDESS dataset has one label per video, and as our Auditory channel processes 1s of speech signal at a time, we perform a voting scheme to obtain one label per video. As our Visual channel processes 1 frame at a time, and our Auditory channel processes 1s of speech signals, we perform an extra step for processing the OMG-Emotion dataset. For each 1s of video, we pair each frame with the entire 1s of audio. This results in each second of video producing 25 pairs of training data. During the evaluation, we perform voting between the model's output of the 25 samples to produce one label per second.

The GWR learns how to create prototype neurons based on the input data distribution, and thus, it is important to measure how well the input data is being represented. In this regard, we calculate the Kullback-Leibler divergence between the normalized distribution of the input data labels and the distribution of the labels associated with the prototype neurons in each of the proposed mechanisms. Analyzing this measure together with the objective performance allows us to evaluate how the prototype neurons are generalizing the input data.

3.2.1 Exp 1.1: The Quality of the Emotion Expression Representation

The OMG-Emotion dataset allows us to evaluate our *AffMem* framework on recognizing continuous emotional expressions. The framework, however, relies heavily on the *CCCNN* representations. Thus, we must guarantee that the learned convolutional filters are robust enough to recognize emotional expressions. The OMG-Emotion dataset, however, does not contain enough individual data samples for the *CCCNN* to learn such robust filters. Thus, we pre-train each individual channel of the CCCNN with the Affectnet and the RAVDESS dataset respectively. We report all the ablation and validation studies for the individual channels on Appendix A.

Once with the channels pre-trained, we fine-tune the entire model using the OMG-Emotion train subset and evaluate it using the test subset. Also in order to demonstrate the importance of the pre-training, we provide an experiment where we train the entire *CCCNN* trained from the scratch using the train subset of the OMG-Emotion and evaluate it with the test subset.

3.2.2 Exp 1.2: Performance of the Prototype Neurons of the General-GWR

As soon as the *CCCNN* is trained with the OMG-Emotion train subset, and its convolutional filters are able to describe emotion expressions, we use the output of its first hidden layer as input to the *General-GWR*. As the *General-GWR* creates prototype neurons which represent the input data, it does not require a massive amount of training samples to learn meaningful representations. To evaluate this assumption, we measure the performance of the *General-GWR* in recognizing multimodal emotion expressions when trained with the train subset of the OMG-Emotion dataset.

Our new *General-GWR* implements a Growing-When-Required network instead of a self-organizing map, as in our previous work [36]. We also proposed the implementation of gamma connections to embed the prototype neurons with a temporal context of the perceived emotion expressions. To assess the impact of these two design decisions and investigate whether they indeed improve emotion recognition, we also provide experiments using a SOM and a common GWR without the gamma memory. For the sake of a fair comparison, all these models were optimized using the TPE optimizer to maximize the CCC on arousal and valence.

3.2.3 Exp 1.3: The Effect of the Dynamic Learning of the Affective Memories

By pre-training the *General-GWR* on the OMG-Emotion train set, we obtained a general emotion recognition mechanism. To obtain a personalized expression modeling, however, we associate the labels coming from the *General-GWR* with the feature vector extracted by the *CCCNN* to train the *affective memories*.

The neurons created by the *affective memories* would be more specific for one particular subject, and thus, we assume that they should provide a better emotion recognition performance in the long run. We evaluate the *affective memories* by performing an online learning experiment where each video of the test subset of the OMG-Emotion dataset, representing one unique subject, is processed by one *affective memory*.

To avoid those expressions with extreme arousal and valence values, we proposed a novel update rule using the neural activation (see Eqs. (8) and (9)) as a learning modulator. To objectively assess its impact on the recognition of the expressions, we perform experiments with and without the proposed update rule.

To obtain an objective averaged continuous measure between the model with and without the modulation, we represent each test video in a percentage scale, where 0 percent is the beginning of the video, and 100 percent is the end of the video. We calculate the averaged CCC over all the test videos between the BMUs and their five neighbors, and the real labels, for each 10 percent of the videos. This way we can observe how the *affective memories* behave when the videos unfold.

3.2.4 Exp 2: Continuous Emotion Recognition

The *AffMem* framework integrates all the previously mentioned mechanisms with the Mood to provide a reliable emotion recognition system for continuous emotion expressions. To evaluate the entire framework, we run an online learning experiment with the OMG-Emotion test subset. The experiment is performed in the same way as *Exp 1.3*, but with the integration of the Mood. The CCC is calculated between the arousal and valence read from the Mood neurons and the original labels. We provide the performance over time and compare the readings from the Mood with the labels from the *General-GWR*, and the *affective memories*, in order to highlight the differences in the recognition dynamics. We also provide a visualization on how the neurons of Mood are formed over time, and how they differ from the ones in the *affective memories*.

TABLE 3
Summary of All Our Experiments: The Investigated Models, Training, Tuning, and Evaluation Datasets (AFF: AffectNet, RAV:RAVDESS, OMG:OMG-EMotion)

Exp.	Model	Datasets		
		Train	Tuning	Evaluation
1.1	*CCCNN*	OMG	-	OMG
	CCCNN	RAV + AFF	OMG	OMG
1.2	SOM	OMG	-	OMG
	GWR	OMG	-	OMG
	General-GWR	OMG	-	OMG
1.3	*Affec. Mem. modulation*	-	-	OMG
	Affec. Mem. no modulation	-	-	OMG
2	*AffMem*	-	-	OMG

In order to simplify the understanding of our experimental setup, we provide in Table 3 a summary of all our experiments, investigated mechanisms, and evaluated datasets.

3.3 Results

3.3.1 Exp 1.1: The Quality of the Emotion Expression Representation

Our experiments with the CCCNN, pre-trained with the AffectNet and RAVDESS datasets, and tuned with the OMG-Emotion train subset reached a CCC of 0.34 for arousal and 0.45 for valence, as reported in Table 4. For comparison, we also train the network from scratch with only the OMG-Emotion dataset. The results, however, are much worse than when tuning the CCCNN only with the OMG-Emotion train subset. This is somehow expected, as the OMG-Emotion train subset does not contain enough data samples to train the individual channels properly, and thus, the pre-training with other datasets is very important.

3.3.2 Exp 1.2: Performance of the Prototype Neurons of the General-GWR

The results of our experiments with the *General-GWR* are reported in Table 5. Our *General-GWR* obtained the best performance when compared to the common GWR and the SOM, achieving a CCC of 0.35 for arousal and 0.47 for valence. The SOM, as expected, achieved the worst results in this experiment, as it was topologically limited by its predefined neighbor configuration and, thus, could not develop robust prototype neurons. Also, the *General-GWR* presented a better performance when compared to the *CCCNN* results. This confirms our assumption that the gamma connections enhanced the recognition by learning small transitions of the expressions, instead of the snapshot representation recognition performed by the *CCCNN*.

TABLE 4
Concordance Correlation Coefficient (CCC), for Arousal and Valence When Evaluating the *CCCNN* With the OMG-Emotion Datasets

Model	Datasets			Arousal	Valence
	Train	Tuning	Evaluation		
CCCNN	OMG	-	OMG	0.17	0.33
CCCNN	AFF+RAV	OMG	OMG	**0.34**	**0.45**

TABLE 5
Concordance Correlation Coefficient (CCC), for Arousal and Valence, When Evaluating the *General-GWR*, a Common GWR, and a Self-Organizing Map (SOM), on the Test Set of the OMG-Emotion

Model	Arousal	Valence	Neurons	KL Divergence	
				Arousal	Valence
SOM	0.32	0.38	400	0.29	0.34
GWR	0.34	0.41	914	0.10	0.09
General-GWR	**0.35**	**0.47**	1253	0.12	0.13

We also provide the number of neurons of each model after training, and the Kullback-Leibler (KL) divergence between the labels associated with the prototype neurons and the training set of the OMG-Emotion

The *General-GWR* created more neurons than the common GWR, which is also an impact from the gamma connections. We assume that the increase in the number of neurons was due to the creation of prototypes that represent the transition between the expressions, while the prototype neurons of the common GWR learned single representations of an expression. Analyzing the associated label distribution for each model in Fig. 7, we observe that the prototype neurons of the General-GWR cover the distribution of the original training data better than the common GWR and the SOM. The *General-GWR* provides, however, a higher KL divergence when compared to the common GWR, which indicates that it learned representations which were not a simple copy of the training dataset, but a generalization of the emotion expression concepts. Combined with the better accuracy, we can confirm that the neurons learned by the *General-GWR* are actually much more robust to recognize unknown expressions than the ones from the common GWR.

3.3.3 Exp 1.3: The Effect of the Dynamic Learning of the Affective Memories

After processing all the frames of each video, the *affective memories* reached a better performance when compared to the *General-GWR*: 0.37 CCC for arousal and 0.51 CCC for

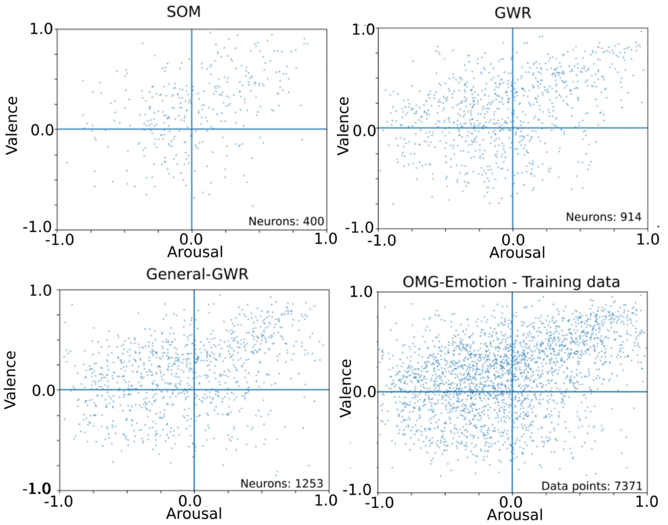

Fig. 7. Arousal and valence distribution of the associated labels of the prototype neurons learned by the SOM, common GWR and *General-GWR* models.

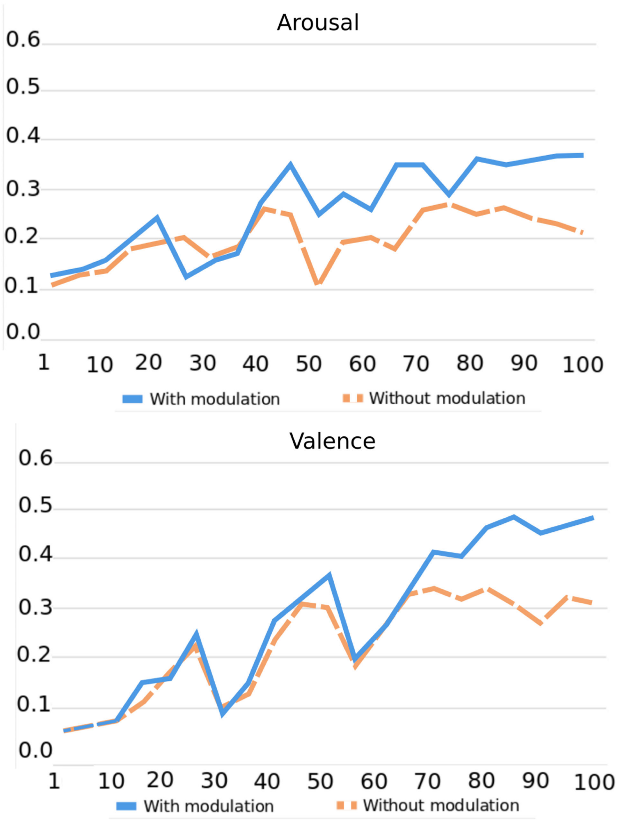

Fig. 8. Average performance, over time, and represented in CCC for arousal and valence of all the videos of the dataset, using the *affective memories* with and without the activation modulation. The videos were represented using a percentage scale (where 100 percent represents the end of all the videos of the dataset) to illustrate the learning behavior of the model.

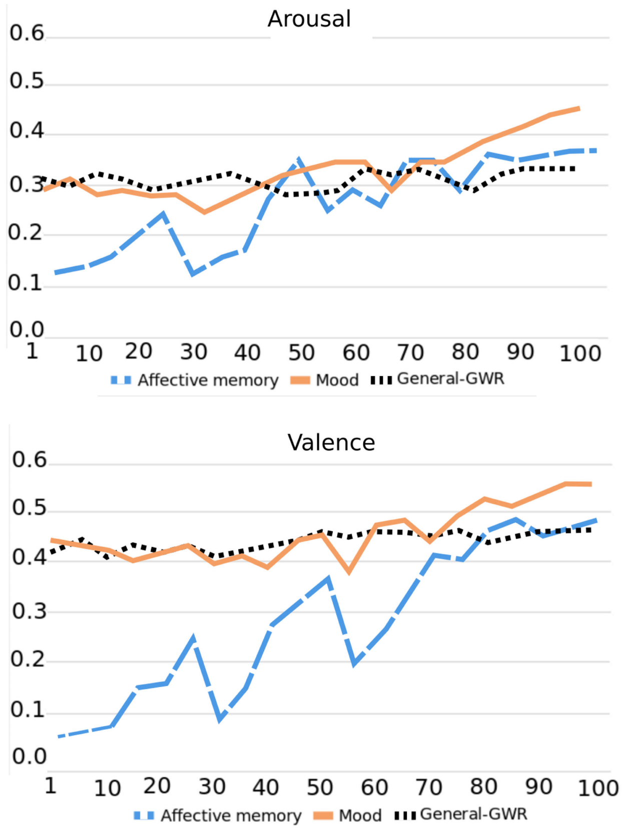

Fig. 9. Average performance, over time, and represented in CCC for arousal and valence of all the videos of the dataset, of the readings from the Mood, the *affective memories*, and the *General-GWR*. The videos were represented using a percentage scale (where 100 percent represents the end of the video) to better illustrate the learning behavior of the model.

valence. At the beginning of the video, however, the performance was very low, reaching 0.12 CCC for arousal and valence. When calculated over the entire video span, the average performance is much worse than the one obtained by the *General-GWR*. When not using the novel update rule, the performance of the *affective memories* was much lower, reaching 0.27 CCC for arousal, and 0.33 CCC for valence at the end of the video. These results are a clear indication that the update rule actually impacts greatly on the formation of robust prototype neurons, as expected.

The most important limitation of the *affective memories*, however, is the time it takes to learn robust representations. Fig. 8 illustrates the performance behavior by plotting the average performance, in CCC for arousal and valence while the video unfolds, for *affective memories* with and without the proposed update rule. We observe that the model with the update rule (see Eq. (6)) presents a continuous improvement on the performance over time, while the model without the update rule stagnates usually in the middle of the videos. The impact of the update rule, as explained in Section 3.3, is to enforce the prototype neurons of the network to have a higher focus on novel information, and thus, learning new expressions faster. This translates into a higher performance after these new expressions have been learned.

The nature of the monologue videos also explains the presence of a recurrent learning pattern on the model with the modulation: a series of performance drops are quickly followed by a performance increase. As the monologue videos present a gradual development in the expressions, they tend to change between known expressions and novel ones. When a novel expression is perceived, the model's performance drops, and after the neurons are quickly created, it increases to a better level than before. As the video unfolds and new expressions are not common anymore, the model is able to maintain a higher performance as it created neurons which represent all the known expressions.

3.3.4 Exp 2: Continuous Emotion Recognition

The performance behavior of our *AffMem* framework for continuous learning is exhibited in Fig. 9. In order to understand better the contributions of the Mood, we plot its performance over time, as the CCC for arousal and valence, together with the performance of the *affective memories* and the *General-GWR*. We observe that the performance of the Mood is influenced by both *General-GWR*, by presenting a constant behavior at the beginning of the videos, and the *affective memories*, which enforce an improvement of the performance over time. The Mood, at the beginning of the videos, maintains a similar performance when compared to the *General-GWR*, while at the end of the video surpasses it, reaching the same performance of the *affective memories* after all the frames were processed. The average performance, per each 1s of the video, of the *AffMem* over the entire videos reach a CCC of 0.45 for arousal, and 0.56 for valence.

TABLE 6
Concordance Correlation Coefficient (CCC), for Arousal and Valence, of Our Final Model and the Current State-of-the-Art Results for the OMG-Emotion Dataset

Model	Arousal	Valence
AffMem	**0.45**	**0.56**
Barros *et al.* [30]	0.43	0.53
Zheng *et al.* [56]	0.35	0.49
Huang *et al.* [12]	0.31	0.45
Peng *et al.* [57]	0.24	0.43
Deng *et al.* [58]	0.27	0.35

The final performance of the *AffMem* framework is better than the current state-of-the-art results on the OMG-Emotion dataset, represented by the winners of the challenge where the dataset was presented [56], [57], [58], as reported in Table 6. All three models also used the pre-training of unisensory convolutional channels to achieve such results, but with neural networks with many parameters to be fine-tuned in end-to-end solutions. The use of attention mechanisms [56] to process the continuous expressions on the videos presented the best results of the challenge, achieving a CCC of 0.35 for arousal and 0.49 for valence. Temporal pooling, implemented as bi-directional Long-Short Time Memories (LSTMs) [57], achieved the second place, with a performance of 0.24 CCC for arousal and 0.43 CCC for valence. The late fusion of facial expressions, speech signals, and text information reached third-best result [58], with a CCC of 0.27 for arousal and 0.35 for valence. Our previous model [30] combined the generalization of generative adversarial networks, which generate conditional facial expressions, to populate a growing-when-required networks in an online manner to achieve 0.35 CCC for arousal and 0.53 for valence based on facial expressions alone. Although this hybrid network outperforms the performance of some multimodal solutions, it required extensive pre-training to allow the adversarial network to converge. The complex attention-based network proposed by Huang *et al.* [12] was able to achieve a CCC of 0.31 in arousal and 0.45 in valence, using only visual information.

4 Discussions

Our *AffMem* model presented state-of-the-art results when recognizing the continuous emotion expressions from the videos of the OMG-Emotion dataset. Our model relies on a combination of the strong pre-training of individual convolutional channels, the gamma connections and the dynamics of the Growing-When-Required networks to depict the temporal information of the expressions, which, performance-wise, was more robust than any of the state-of-the-art models. The interplay between the general and personalized perception was not addressed by any of the current models to recognize the OMG-Emotion dataset, which could explain our better performance.

Also, it is important to notice that our proposed framework is modular, and any of its neural components can be replaced when necessary. If a better feature extractor would be available in the future, the CCCNN could be easily replaced. Also, on tasks where personalization is not important, the BMUs of the *General-GWR* could be used to represent affect, for example.

In this section, we dissect the behavior of our model by demonstrating how it processes two of the videos of the OMG-Emotion dataset, which helps us to explain how the model deals with the gradual changes in the emotion expressions better than the recent end-to-end neural networks.

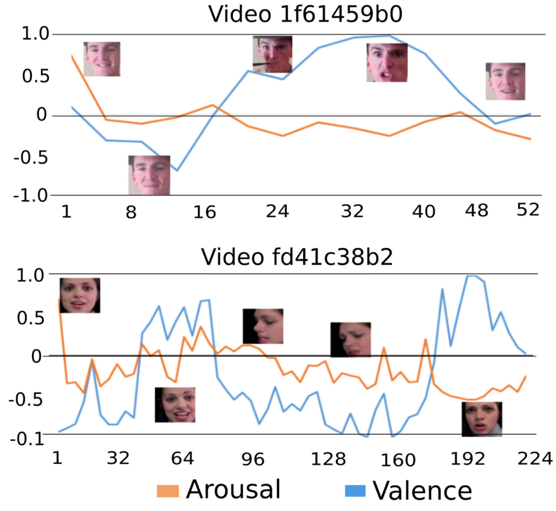

Fig. 10. Arousal and valence labels, and facial expression examples, over time (in seconds) of the longest (fd41c38b2 - 224s) and shortest (1f61459b0 - 52s) videos of the OMG-Emotion dataset.

4.1 Recognizing Emotions in Monologues

In recent years, several "in-the-wild" datasets were published, mostly based on web-crawlers to obtain emotion expression samples from different sources over the internet. This impacted heavily on the development of emotion expression recognition models, mostly based on extreme generalization. Such models presented, however, problems when applied to real-world scenarios. One of the reasons lies directly on their data-driven learning strategies: although such models are focused on the generalization of the expressions, they neglect the gradual and natural transitions between expressed emotional states. Even very popular "in-the-wild" datasets, such as the Emotions-in-The-Wild [59] and the Affect-in-the-wild [60] challenge dataset present expressions which are self-contained, and thus, difficult to be represented in continuous interactions.

The monologues of the OMG-Emotion, however, have an important characteristic: by nature, the expressions present in them are gradually changing over time, based on a contextual development. The persons performing the monologues use emotions as a mean to tell a story or to interpret a certain situation. As the videos are usually longer than the ones present in common "in-the-wild" datasets, these emotion changes are much more impacted by the natural transition between emotional states. Translating this behavior into computational models becomes difficult if such models are designed to achieve maximum generalization on instantaneous expression recognition. The impact of personalization, in the case of the monologues, also play a substantial role, as the person performing the monologue adds his or her own characteristics in the expressions. As a simple example to illustrate the affective state transitions, we present the plot of ground truth annotation over time of the shortest (video 1f61459b0), with 52 seconds, and the longest (video fd41c38b2), with 224 seconds, of the videos in the OMG-Emotion test set in Fig. 10.

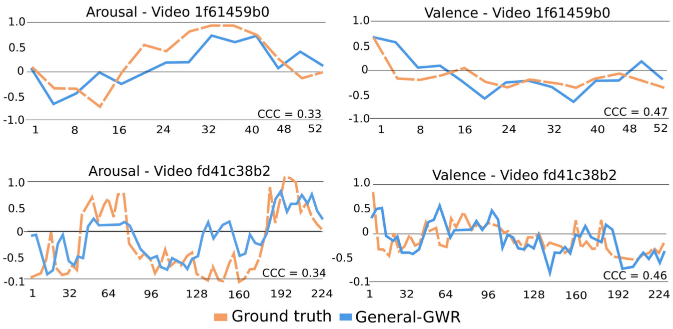

Fig. 11. Arousal and Valence output of the *General-GWR* and ground truth label, over time (in seconds) of the longest (fd41c38b2 - 224s) and shortest (1f61459b0 - 52s) videos of the OMG-Emotion dataset.

4.2 The Effect of Generalization

When processing the two videos represented in Fig. 10, the *General-GWR* presents a performance, measured in CCC, of 0.33 for arousal and 0.47 for valence for the shortest video, and 0.34 for arousal and 0.46 for valence for the longest video. When plotting the recognition of the *General-GWR* over time together with the videos' ground-truth, illustrated in Fig. 11, we observe that the temporal processing of the GWR adds a certain delay on emotion recognition. The same affective transitions are there but slightly shifted over time. This happens mostly due to the neurons on the *General-GWR* having a context with depth 3, meaning the BMU calculation will be heavily impacted by the three last seconds of the expression. This is beneficial for the *General-GWR*, as it maintains the same recognition trend as the ground-truth, and thus, achieves a steady Concordance Correlation Coefficient, but it limits the maximum performance that the *General-GWR* can achieve. We assume that the same behavior happens with the deep neural model which relies on temporal processing proposed by the winners of the OMG-Emotion dataset challenge [56]. The transitions between affective states happen gradually, but with a sharp change of directions. Possibly this is why the performance of the model proposed by Zheng *et al.* [56] was similar to the *General-GWR*.

4.3 The Effect of Personalization

When training the *affective memories* with the same two videos, we clearly observe how it adapts over time. Fig. 12 illustrates the plot with the arousal and valence of the *affective memories*, the ground truth labels, and the network activation over time. The *affective memories* are heavily impacted by drastic changes on the arousal and valence, so when these drastic changes do not occur, the network does not learn much. This effect can be easily observed by tracking the network's activation over time. When the activation decreases, the network adds more neurons and the neural update is stronger, meaning it is rapidly adapting.

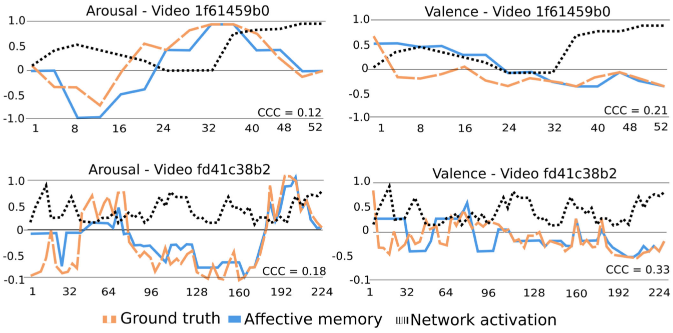

Fig. 12. Arousal and Valence output of the *affective memories*, ground truth label, and network activation over time (in seconds) of the longest (fd41c38b2 - 224s) and shortest (1f61459b0 - 52s) videos of the OMG-Emotion dataset.

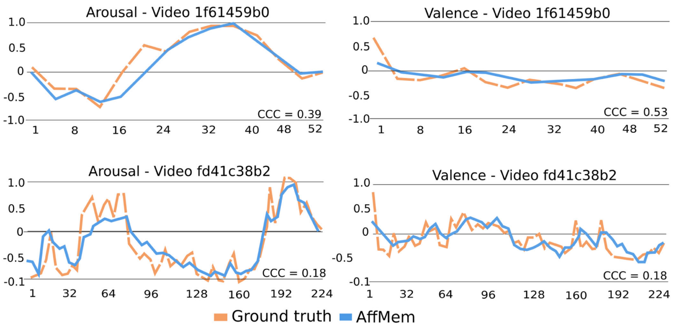

Fig. 13. Arousal and Valence output, over time (in seconds), from the *AffMem* compared to the ground truth of the longest (fd41c38b2 - 224s) and shortest (1f61459b0 - 52s) videos of the OMG-Emotion dataset.

By the end of the videos, the *affective memories* experienced most of the possible expressions that the monologue videos contained, and the predicted arousal and valence are much more accurate when compared with the ground truth. This effect was demonstrated as the increase of the activation of the neurons. The delay effect, although still present, is much smaller when compared with the *General-GWR*, due to the network learning only the specific neurons which are necessary to recognize that specific person. Thus, the transitions learned by the *affective memories* are much more sparse, which is represented by the number of neurons created for each video: 100 neurons for the shortest video, and 400 neurons for the longest one.

As the *affective memories* perform better with more samples to learn from, they present a better overall performance when recognizing longer videos. This is illustrated by the better performance on the longest video, reaching a CCC of 0.21 for arousal and 0.33 for valence, compared to the shortest video, where they reach a CCC of 0.12 for arousal and 0.18 for valence.

4.4 The Interplay Between Personalization and Generalization

Our experiments show that the *AffMem* model presents the best overall performance on recognizing expressions on the OMG-Emotion monologue videos, but the online learning and adaptation of these expressions is the most important contribution of this research. Different from purely batch-learning models, our model adapts to the expressed emotions over time, which showed to be very important to achieve good performance. The *AffMem* clearly combines the characteristics of the general emotion perception from the *General-GWR*, and the personalization of the *affective memories*, into one continuous learning framework. This was demonstrated by plotting the average performance of the model over time, as illustrated in our experiments in Fig. 9, but also when plotting, in Fig. 13, the ground truth labels and the output of the

AffMem overtime for the shortest and the longest videos of the OMG-Emotion test set.

We observe that our framework, although not modeling the ground truth of both videos perfectly, does depict the behavioral change between the transitions of the emotional expressions better than the *General-GWR* and the *affective memories*. In the longest video, we can see the modulation of the *affective memories* to be much stronger than on the shortest video, as the arousal and valence follow the same pattern as the *General-GWR* at the beginning of the video and the same pattern of the *affective memories* by the end of it.

4.5 Arousal Versus Valence Recognition

Another fact we observed in our model was the lower performance of arousal when compared with valence. To recognize arousal is a more difficult challenge when compared to valence recognition, as shown by our results and the state-of-the-art models. This could be due to the following two reasons. First, in humans, the persistence of valence perception is longer than arousal [61], which could indicate a certain bias on the gradual annotations of the videos. Second, arousal is harder to depict via facial expressions, as it is more salient when expressed via body movements [62].

5 Conclusion

Emotions are present in many aspects of our lives, which includes interpersonal communication, learning and experiencing new things, and remembering past events. In particular, embedding concepts and ideas related to emotions into intelligent systems will improve greatly their capability to act as active social agents, through the processing, understanding, and synthesis of social behavior. However, there are still great challenges to be solved before this can happen. One of them is the inability of current emotion expression recognition systems, mostly based on end-to-end deep neural models, to adapt quickly to novel information.

We proposed here the Affective Memory framework which is composed of three self-organizing mechanisms based on Growing-When-Required networks: the *General-GWR*, responsible to provide a general emotion perception, the *affective memories*, which learn online, specific characteristics of how a person expresses emotions, and the *Mood*, which generates an arousal and valence reading of continuous expressions.

We evaluate our model with a set of different ablation studies to investigate the contributions of each of these mechanisms. Finally, we used the OMG-Emotion dataset to evaluate the entire framework on a continuous emotion expression scenario. Our model presented a state-of-the-art performance for the recognition of both arousal and valence. Furthermore, we discussed that the proposed mechanisms of the *AffMem* are appropriate for processing the monologue videos of the OMG-Emotion dataset, and explain why our model would be suitable for real-world scenarios.

Our model deals with multimodal expressions through the simultaneous processing of facial expressions and the speech signal. This, however, limits the model's applicability where these modalities may not be perceived reliably at the same time. An asynchronous processing of multimodal perception would benefit the model in these cases. The incorporation of different general perception modalities, such as language, touch, or even specific modalities such as a conversation context, or intrinsic motivation, would also be encouraged for future work.

Appendix A

CCCNN Ablation Experiments and Results

Our assumption is that to make the *CCCNN* able to represent emotion expressions in a reliable manner, we must pre-train the individual channels using the AffectNet and the RAVDESS dataset, and fine-tune the entire network with the OMG-Emotion dataset.

To evaluate such an assumption, we must, first, assess the impact that the individual channels have on each specific dataset and on the OMG-Emotion dataset. To evaluate the individual channels, we train and evaluate the Face channel and the Auditory channel with the individual training and testing protocols of the AffectNet and with the RAVDESS respectively. Then, we fine-tune each channel using the train subset of the OMG-Emotion dataset and repeat the evaluation on the test subset of the OMG-Emotion. Finally, to obtain a better understanding of the impact of the pre-training, we train each individual channel from scratch with the training subset of the OMG-Emotion and evaluate them using the test subset.

The Face channel is evaluated using the CCC between the predicted samples and the true labels for arousal and valence of the AffectNet dataset. When evaluating the Auditory channel with the RAVDESS dataset, we perform a leave-one-actor-out experiment and compute the mean accuracy between the channel's output and the true labels of each video. Evaluation on the OMG-Emotion dataset is done by calculating the CCC for every second of the videos.

A.1 Experiment Summary

In order to simplify the understanding of our CCCNN ablation experiment setup, we provide in Table 7 a summary of all our investigated mechanisms and evaluated datasets.

TABLE 7
Summary of All The Ablation Studies and Results Using the CCCNN: The Investigated Models, Training, Tuning, and Evaluation Datasets (AFF: AffectNet, RAV: RAVDESS, OMG: OMG-EMotion)

Model	Datasets		
	Train	Tuning	Evaluation
Face c.	Aff	-	Aff
Face c.	OMG	-	OMG
Face c.	Aff	OMG	OMG
Auditory c.	RAV	-	RAV
Auditory c.	OMG	-	OMG
Auditory c.	RAV	OMG	OMG

A.2 Results

We first investigated the performance of the individual channels of the CCCNN and report them in Table 8 together with the results of the current state-of-the-art models for each evaluated dataset. The Auditory channel presents a

TABLE 8
Concordance Correlation Coefficient (CCC), for Arousal and Valence, and the Categorical Accuracy When Evaluating the Individual Channels With the AffectNet and the RAVDESS Datasets

RAVDESS		
Model	Accuracy	
Auditory channel	70.5	
VGG16 [63]	**71.0**	
ResidualNet [64]	64.8	
Wavelet transformation [65]	60.1	
AffectNet		
Model	Arousal	Valence
Face channel	**0.46**	**0.65**
AlexNet [66]	0.34	0.60

competitive result on the speech signals of the RAVDESS dataset. Our model, however, was a light-weight neural network, and yet performed similarly to the pre-trained VGG16 [63], which has much more training parameters, and it was better than a Residual Network implementation [64], which contains recurrent connections which are more complex to train. Both the pre-trained VGG16 and the Residual Network implementation require a strong pre-training with other datasets in order to achieve good performance on the RAVDESS data. The Auditory channel also performs better than solutions which use standard signal-processing models to represent the audio and rely upon heavily tuned SVMs for the recognition [65].

To the best of our knowledge, the only reported CCC for arousal and valence on the AffectNet corpus comes from the authors of the dataset themselves [14]. This is probably the case as most of the research uses the AffectNet dataset to pre-train neural models for generalization tasks in other datasets, without reporting the performance on the AffectNet itself. The baseline provided by the authors uses an AlexNet convolutional neural network [66] re-trained to recognize arousal and valence. Our Face channel provided better performance, improving the CCC by more than 0.12 for arousal and 0.05 for valence.

The results of the training and evaluation with the OMG-Emotion dataset can be found in Table 9. Both channels improve drastically the performance on recognition arousal and valence on the OMG-Emotion test subset when pre-trained with the AffectNet and RAVDESS datasets.

TABLE 9
Concordance Correlation Coefficient (CCC), for Arousal and Valence, When Evaluating the Individual Channels With the OMG-Emotion Datasets

Model	Datasets			Arousal	Valence
	Train	Tuning	Evaluation		
Auditory c.	OMG	-	OMG	0.09	0.21
Face c.	OMG	-	OMG	0.13	0.27
Auditory c.	RAV	OMG	OMG	0.21	0.30
Face c.	Aff	OMG	OMG	0.24	0.44

ACKNOWLEDGMENTS

This work was partially supported by the German Research Foundation (DFG) under project CML (TRR 169). The authors would like to thank German I. Parisi for important suggestions and support on the development of our model, and Katja Koesters for the review of this manuscript.

REFERENCES

[1] P. Ekman and W. V. Friesen, “Constants across cultures in the face and emotion,” *J. Pers. Soc. Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.

[2] F. Cavallo, F. Semeraro, L. Fiorini, G. Magyar, P. Sinčák, and P. Dario, “Emotion modelling for social robotics applications: A review,” *J. Bionic Eng.*, vol. 15, no. 2, pp. 185–203, 2018.

[3] S. Hamann and T. Canli, “Individual differences in emotion processing,” *Curr. Opinion Neurobiol.*, vol. 14, no. 2, pp. 233–238, 2004.

[4] U. Hess, C. Blaison, and K. Kafetsios, “Judging facial emotion expressions in context: The influence of culture and self-construal orientation,” *J. Nonverbal Behav.*, vol. 40, no. 1, pp. 55–64, 2016.

[5] P. E. Griffiths, “III. Basic emotions, complex emotions, machiavellian emotions 1,” *Roy. Inst. Philosophy Supplements*, vol. 52, pp. 39–67, 2003.

[6] L. F. Barrett, “Solving the emotion paradox: Categorization and the experience of emotion,” *Pers. Soc. Psychol. Rev.*, vol. 10, no. 1, pp. 20–46, 2006.

[7] S. Afzal and P. Robinson, “Natural affect data - collection and annotation in a learning context,” in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interaction*, 2009, pp. 1–7.

[8] D. Mehta, M. Siddiqui, and A. Javaid, “Facial emotion recognition: A survey and real-world user experiences in mixed reality,” *Sensors*, vol. 18, no. 2, 2018, Art. no. 416.

[9] M. Soleymani, M. Pantic, and T. Pun, “Multimodal emotion recognition in response to videos,” *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 211–223, Apr.–Jun. 2012.

[10] M. E. Kret, K. Roelofs, J. J. Stekelenburg, and B. de Gelder, “Emotional signals from faces, bodies and scenes influence observers’ face expressions, fixations and pupil-size,” *Front. Hum. Neurosci.*, vol. 7, 2013, Art. no. 810.

[11] D. Hazarika, S. Gorantla, S. Poria, and R. Zimmermann, “Self-attentive feature-level fusion for multimodal emotion detection,” in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval*, 2018, pp. 196–201.

[12] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-H. Chen, “Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 5866–5870.

[13] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, “Collecting large, richly annotated facial-expression databases from movies,” *IEEE Multimedia*, vol. 19, no. 3, pp. 34–41, Thirdquarter 2012.

[14] 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.*, vol. 10, no. 1, pp. 18–31, 2017.

[15] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, “Aff-wild: Valence and arousal ‘in-the-wild’challenge,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1980–1987.

[16] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, “Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph,” in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, vol. 1, pp. 2236–2246.

[17] E. Marinoiu, M. Zanfir, V. Olaru, and C. Sminchisescu, “3D human sensing, action and emotion recognition in robot assisted therapy of children with autism,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2158–2167.

[18] J. Yang, K. Wang, X. Peng, and Y. Qiao, “Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction,” in *Proc. Int. Conf. Multimodal Interaction*, 2018, pp. 594–598.

[19] W. Y. Choi, K. Y. Song, and C. W. Lee, “Convolutional attention networks for multimodal emotion recognition from speech and text data,” in *Proc. Grand Challenge Workshop Hum. Multimodal Lang.*, 2018, pp. 28–34.

[20] Z. Du, S. Wu, D. Huang, W. Li, and Y. Wang, “Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition,” in *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2019.2940224.

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

[22] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, and Y. Zong, "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, vol. 309, pp. 27–35, 2018.

[23] S. Costa, A. Brunete, B.-C. Bae, and N. Mavridis, "Emotional storytelling using virtual and robotic agents," *Int. J. Humanoid Robot.*, vol. 15, no. 03, 2018, Art. no. 1850006.

[24] C. Lytridis, E. Vrochidou, and V. Kaburlasos, "Emotional speech recognition toward modulating the behavior of a social robot," in *Proc. JSME Annu. Conf. Robot. Mechatronics*, 2018, pp. 1A1–B14.

[25] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," in *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2018.2874986.

[26] P. Barros, N. Churamani, E. Lakomkin, H. Sequeira, A. Sutherland, and S. Wermter, "The OMG-emotion behavior dataset," in *Proc. Int. Joint Conf. Neural Netw.*, 2018, pp. 1408–1414.

[27] S. Berretti, M. Daoudi, P. Turaga, and A. Basu, "Representation, analysis, and recognition of 3D humans: A survey," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 1s, 2018, Art. no. 16.

[28] Y. Kim, B. Yoo, Y. Kwak, C. Choi, and J. Kim, "Deep generative-contrastive networks for facial expression recognition," 2017, *arXiv: 1703.07140*.

[29] S. Saha, R. Navarathna, L. Helminger, and R. M. Weber, "Unsupervised deep representations for learning audience facial behaviors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 1132–1137.

[30] P. Barros, G. Parisi, and S. Wermter, "A personalized affective memory model for improving emotion recognition," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 485–494.

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

[32] H. Ding, S. K. Zhou, and R. Chellappa, "FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 118–126.

[33] G. Pons and D. Masip, "Supervised committee of convolutional neural networks in automated facial expression analysis," *IEEE Trans. Affect. Comput.*, vol. 9, no. 3, pp. 343–350, Thirdquarter 2018.

[34] G. Zen, E. Sangineto, E. Ricci, and N. Sebe, "Unsupervised domain adaptation for personalized facial emotion recognition," in *Proc. 16th Int. Conf. Multimodal Interaction*, 2014, pp. 128–135.

[35] S. E. Kahou *et al.*, "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[36] P. Barros and S. Wermter, "Developing crossmodal expression recognition based on a deep neural model," *Adaptive Behav.*, vol. 24, no. 5, pp. 373–396, 2016.

[37] P. Barros and S. Wermter, "A self-organizing model for affective memory," in *Proc. Int. Joint Conf. Neural Netw.*, 2017, pp. 31–38.

[38] J. D. Mayer, M. DiPaolo, and P. Salovey, "Perceiving affective content in ambiguous visual stimuli: A component of emotional intelligence," *J. Pers. Assessment*, vol. 54, no. 3/4, pp. 772–781, 1990.

[39] J. A. Russell, "Cross-cultural similarities and differences in affective processing and expression," in *Emotions and Affect in Human Factors and Human-Computer Interaction*. Amsterdam, The Netherlands: Elsevier, 2017, pp. 123–141.

[40] R. Sprengelmeyer, M. Rausch, U. T. Eysel, and H. Przuntek, "Neural structures associated with recognition of facial expressions of basic emotions," *Proc. Roy. Soc. London B: Biol. Sci.*, vol. 265, no. 1409, pp. 1927–1931, 1998.

[41] G. I. Parisi, J. Tani, C. Weber, and S. Wermter, "Lifelong learning of human actions with deep neural network self-organization," *Neural Netw.*, vol. 96, pp. 137–149, 2017.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[43] Y. Fregnac, C. Monier, F. Chavane, P. Baudot, and L. Graham, "Shunting inhibition, a silent step in visual cortical computation," *J. Physiol.*, vol. 97, pp. 441–451, 2003.

[44] 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, Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[45] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 2546–2554.

[46] A. Soltoggio, K. O. Stanley, and S. Risi, "Born to learn: The inspiration, progress, and future of evolved plastic artificial neural networks," *Neural Netw.*, vol. 108, pp. 48–67, 2018.

[47] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Netw.*, vol. 113, pp. 54–71, 2019.

[48] D. J. Mankowitz *et al.*, "Unicorn: Continual learning with a universal, off-policy agent," 2018, *arXiv: 1802.08294*.

[49] B. De Vries and J. C. Principe, "The gamma model–A new neural model for temporal processing," *Neural Netw.*, vol. 5, no. 4, pp. 565–576, 1992.

[50] S. Marsland, J. Shapiro, and U. Nehmzow, "A self-organising network that grows when required," *Neural Netw.*, vol. 15, no. 8, pp. 1041–1058, 2002.

[51] D. Maltoni and V. Lomonaco, "Continuous learning in single-incremental-task scenarios," *Neural Netw.*, vol. 116, pp. 56–73, 2019.

[52] R. F. de Mello, Y. Vaz, C. H. Grossi, and A. Bifet, "On learning guarantees to unsupervised concept drift detection on data streams," *Expert Syst. Appl.*, vol. 117, pp. 90–102, 2019.

[53] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS One*, vol. 13, no. 5, 2018, Art. no. e0196391.

[54] B. Amos *et al.*, "OpenFace: A general-purpose face recognition library with mobile applications," Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-CS-16–118, 2016.

[55] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, pp. 255–268, 1989.

[56] Z. Zheng, C. Cao, X. Chen, and G. Xu, "Multimodal emotion recognition for one-minute-gradual emotion challenge," 2018, *arXiv: 1805.01060*.

[57] S. Peng, L. Zhang, Y. Ban, M. Fang, and S. Winkler, "A deep network for arousal-valence emotion prediction with acoustic-visual cues," 2018, *arXiv: 1805.00638*.

[58] D. Deng, Y. Zhou, J. Pi, and B. E. Shi, "Multimodal utterance-level affect analysis using visual, audio and text features," 2018, *arXiv: 1805.00625*.

[59] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion recognition in the wild challenge 2013," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2013, pp. 509–516.

[60] D. Kollias *et al.*,"Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *Int. J. Comput. Vis.*, vol. 127, pp. 907–929, 2019.

[61] P. Gomez, P. Zimmermann, S. Guttormsen Sch är, and B. Danuser, "Valence lasts longer than arousal: Persistence of induced moods as assessed by psychophysiological measures," *J. Psychophysiol.*, vol. 23, no. 1, pp. 7–17, 2009.

[62] M. V. Peelen, A. P. Atkinson, F. Andersson, and P. Vuilleumier, "Emotional modulation of body-selective visual areas," *Soc. Cogn. Affect. Neurosci.*, vol. 2, no. 4, pp. 274–283, 2007.

[63] A. S. Popova, A. G. Rassadin, and A. A. Ponomarenko, "Emotion recognition in sound," in *Proc. Int. Conf. Neuroinformatics*, 2017, pp. 117–124.

[64] Y. Zeng, H. Mao, D. Peng, and Z. Yi, "Spectrogram based multi-task audio classification," *Multimedia Tools Appl.*, vol. 78, pp. 3705–3722, 2019.

[65] P. Shegokar and P. Sircar, "Continuous wavelet transform based speech emotion recognition," in *Proc. 10th Int. Conf. Signal Process. Commun. Syst.*, 2016, pp. 1–8.

[66] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

Pablo Barros received the BSc degree in information systems from the Universidade Federal Rural de Pernambuco, Brazil, the MSc degree in computer engineering from the Universidade de Pernambuco, Brazil, and the PhD degree in computer science from Universitt Hamburg, Germany, Currently, he is a research scientist at the Italian Institute of Technology, Italy. His main research interests include artificial neural networks and its applications to affective and social robots. He has been a guest editor of the journals the *"IEEE Transactions on Affective Computing"*, *"Frontiers on Neurorobotics"*, and *"Elsevier Cognitive Systems Research"*. He took part in the proposition and organization committee, as publication chair, of the Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics 2020.

Emilia Barakova Her main research interests are in human-robot interaction, cognitive robotics, interaction design and social signal processing. She is a head of TU/e Social Robotics Lab and an assistant professor in Embodied Social Agents for social inclusion at the Department of Industrial Design of the Eindhoven University of Technology, The Netherlands. She has held research positions at RIKEN Brain Science Institute, Japan, GMD-Japan Research Lab, the University of Groningen, The Netherlands, and Bulgarian Academy of Science, Bulgaria. She is an editor of the *Personal and Ubiquitous Computing Journal* (Springer, IF 1,92) and the *International Journal of Social Robotics*, (Springer, IF 2.01). She has organized several IEEE and ACM conferences.

Stefan Wermter is full professor with the University of Hamburg, Germany, and director of the Knowledge Technology Research Institute. His main research interests are in the fields of neural networks, hybrid knowledge technology, neuroscience-inspired computing, cognitive robotics, and human-robot interaction. He is has been an associate editor of the journals the *'Transactions on Neural Networks and Learning Systems'*, and is an associate editor of the *'Connection Science'*, and the *'International Journal for Hybrid Intelligent Systems'* and he is on the editorial board of the journals *'Cognitive Systems Research'*, *'Cognitive Computation'*, and *'Journal of Computational Intelligence'*. Currently, he serves as co-coordinator of the international collaborative research centre on Crossmodal Learning (TRR-169) and is the coordinator of the European Training Network SECURE on safety for cognitive robots. In 2019 he has been elected as the president of the European Neural Network Society until 2022.