A Self-Organizing Model for Affective Memory

Pablo Barros
Department of Computer Science
University of Hamburg
Hamburg, Germany
barros@informatik.uni-hamburg.de

Stefan Wermter
Department of Computer Science
University of Hamburg
Hamburg, Germany
wermter@informatik.uni-hamburg.de

Abstract—**Emotions are related to many different parts of our lives: from the perception of the environment around us to different learning processes and natural communication. Therefore, it is very hard to achieve an automatic emotion recognition system which is adaptable enough to be used in real-world scenarios. This paper proposes the use of a growing and self-organizing affective memory architecture to improve the adaptability of the Cross-channel Convolution Neural Network emotion recognition model. The architecture we propose, besides being adaptable to new subjects and scenarios also presents means to perceive and model human behavior in an unsupervised fashion enabling it to deal with never seen emotion expressions. We demonstrate in our experiments that the proposed model is competitive compared with the state-of-the-art approach, and how it can be used in different affective behavior analysis scenarios.**

I. Introduction

One of the most desired characteristics of autonomous computational models is the ability to respond to social interactions, mostly through the perception and expression of affective behavior [1]. Giving a computational system the capability to perceive affective components on different experiences would change how it interacts with persons in a particular scenario [2]. A simple example is a robot which is capable of determining the affective state of persons in a room and use an estimation as part of its own decision-making process, in a similar way as humans do [3]. This capability could be part of a higher-level cognition process, which can enhance the interaction skill, or even create moral discernment about the information that the robot is receiving.

The first step towards using emotion understanding in autonomous systems is to give them the capability to perceive emotional behavior. In this sense, the area of affective computing was introduced in the 1990s [4] and is presents different solutions to introduce emotion concepts in computational systems. Early works in this field show solutions on emotion perception, mostly in the sense of creating universal descriptors for visual [5] and auditory streams [6]. Most of this work was based on the psychological research of Ekman et al. [7], which introduced the concept of universal emotions. In their study, they showed that there are six emotions which can be understood by any person, independent of the person's age, gender or cultural background: "Disgust", "Fear", "Happiness", "Surprise", "Sadness" and "Anger".

Given the complexity of human emotional behavior, the early solutions on affective computing were usually restricted to a certain domain, such as static images, simple backgrounds or lighting conditions and could not be generalized to a natural communication scenario. This can be explained by the fact that, in a natural communication, a person will show a spontaneous behavior which can aggregate more than one of the characteristics of the universal emotions but also some unknown behavior, increasing the complexity of the task [8].

Recent research in affective computing approaches the topic from different perspectives, approximating their solutions to how the human brain processes and interprets emotions [9], [10]. These models can usually deal with complex tasks, such as recognizing emotion expressions from multimodal streams [11], and even spontaneous expressions [12], [13]. However, such models still have strong constraints: they suffer from poor generalization, need high computational resources or cannot adapt to different scenarios or situations.

The human brain has the capability to associate emotion characteristics with different persons, objects, places and experiences [14]. This capability helps us to perceive different affective behavior better from others, but also increases our capability to adapt and learn different emotion states. These affective characteristics of memories are used by our brain to identify for example when a person is lying, or when a situation is dangerous, and to generalize this knowledge to new persons or situations [15]. Together with the ability to identify spontaneous expressions, the use of such affective memories can help us to increase our perception of different emotional behavior [16].

One of the most common constraints on recent work on affective computing is a restriction to learn and adapt to new information. In previous work [17], we introduced the use of self-organizing networks to increase the learning capability of our deep learning emotion recognition system. Our system uses the capability of convolutional neural networks to learn how to create a representation of visual-auditory expressions and introduces the ability to learn different emotion concepts with a self-organizing layer. This model showed the capability to recognize spontaneous expressions. However, by using a self-organizing layer with fixed topology, the model has a restrictive memory capacity.

In this paper, we propose an extension to this model, by using Growing-When-Required (GWR) self-organizing networks to expand the capability of the model to learn new emotion concepts. We also present a new model for affective memory,

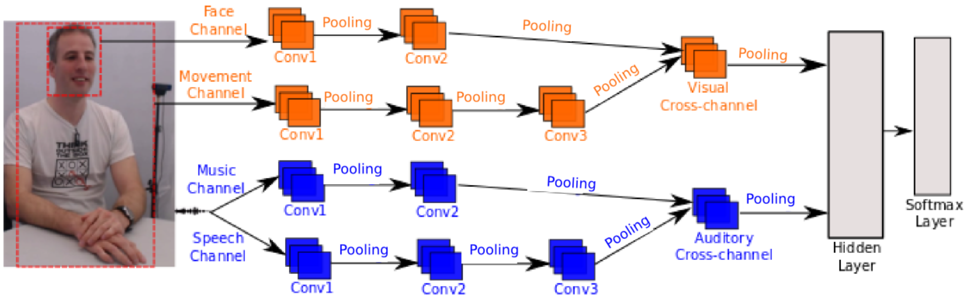

Fig. 1. Final crossmodal architecture, which extracts features from the visual and the auditory input and classifies them in emotion expressions. We connect the outputs of each stream to a fully connected hidden layer and then to a softmax layer, which will give us a classification probability.

which gives our previous model the capability to learn from different subjects, scenarios, and situations in a dynamic way, with very few restrictions.

To evaluate our model we use two corpora: the first one is the dataset for the Emotions-in-The-Wild challenge [18], which contains scenes from different movies. This dataset is used to evaluate the performance of our model in determining spontaneous and natural expressions. We then proceed with the recording of the second corpus with different human-human interactions. Two humans improvise a discussion about a certain topic, and we use our affective memory model to capture different emotion behavior within the topics.

II. Learning crossmodal emotion concepts

To classify spontaneous emotion expressions properly, first, the observation of various different modalities is necessary. Second, the concept of the universal emotions itself is not precise, and the idea of classifying what another person is expressing based on very strict concepts makes the analysis of such models difficult.

Dealing with such a set of restricted emotions is a serious constraint to affective computing systems. Humans have the capability to learn emotion expressions and adapt their internal representation to a newly perceived emotion. This is explained by Hamlin [19] as a developmental learning process. Her work shows that human babies perceive interactions as two very clear directions: positive and negative. When the baby is growing, this perception is shaped based on the observation of human interaction. Eventually, concepts such as the five universal emotions are formed. After observing individual actions toward others, humans can learn how to categorize complex emotions and also concepts such as trust and empathy [20], [21].

To deal with multimodal spontaneous expressions, we make use of the Cross-channel Convolution Neural Network (CC-CNN) and use a self-organizing layer to learn new emotion concepts [17]. In this paper, we extend the self-organizing layer by the means of a Growing-When-Required network (GWR) [22]. This network has the capability to expand and contract its topology, and thus has an unrestricted potential to learn new emotion information.

A. Cross-channel Convolution Neural Network

Spontaneous expressions are composed of multimodal and non-standardized concepts, and thus are very difficult to model in a computational solution. Face expressions, movements, and auditory signals contribute, and have been shown to be necessary [23] for a higher-level understanding of more complex emotion behavior. To be able to represent multimodal spontaneous expressions, we make use of the Cross-channel Convolution Neural Network (CCCNN) architecture. The CCCNN has several channels, each one representing one determined modality and composed of an independent sequence of convolution and pooling layers.

Our implementation is composed of two main streams: a visual and an auditory one. Each stream is divided into two specific channels: a face and a movement one for the visual stream and a music and speech one for the auditory stream. In the last layer of each stream, a Cross-channel creates a high-level representation of the two specific channels, and this is then used as a representation of the multimodal expression.

We connect each Cross-channel with a fully connected hidden layer, with 500 units, which is then connected to a softmax layer. This way, each modality, visual and auditory, has its own high abstraction level representation preserved. Figure 1 illustrates our final architecture.

1) Visual Stream: Our visual stream is inspired by the primate visual cortex [24] with the two channels. The first channel is responsible for learning and extracting information about the facial expression, such as contour, shapes, textures and poses, and it is directly related to the encoding of information found in the ventral area of the primate visual cortex. The second channel learns how to code orientation, direction, and speed of body movements, similar to the information coded by the dorsal area of the primate visual cortex.

Our Face channel receives a cropped face of the person, which is done using the Viola-Jones algorithm. This channel is composed of two convolution and pooling layers, where the first convolution layer implements 5 filters with cubic receptive fields, each one with a dimension of 5x5x3. The second layer implements 5 filters, with a dimension of 5x5, and a shunting inhibitory field. The cubic receptive field allows the model to process sequential information, and the shunting inhibitory field was shown to improve the specification of the features extracted in that layer [17]. Both layers implement 2x2 pooling. Each cropped face is resized to 50x50 pixels, and we use a total of 9 frames to represent an expression.

Our Movement channel receives a movement representation, which is obtained by subtracting 10 consecutive frames and generating one movement representation [25]. This channel implements three convolution and pooling layers, where the first convolution layer implements 5 filters with cubic receptive fields with dimensions of 5x5x3. The second and third channel implement 5 filters, each one with a dimension of 5x5 and all channels implement max-pooling with a receptive field of 2x2. The movement images are resized to a size of 128x96 pixels, and we use a total of 3 motion representations together.

In the last layer of our visual channel, we use a Cross-channel with 10 convolution filters, each one with a dimension of 3x3 and a max-pooling with a receptive field of 2x2. We have to ensure that the input of the Cross-channel has the same dimension, to do so we resize the output representation of the Movement channel to 9x9, the same as the Face channel.

2) Auditory Representation: Our auditory stream is also composed of two channels: a Music channel and a Speech channel. This is also inspired by how the ventral and dorsal areas of the brain process different auditory information [26]. While the ventral stream deals with speech information, the dorsal one maps auditory sensory representation. In earlier stages within the dorsal stream, the auditory information is decomposed into a series of signals which are not connected to phonetic information.

To specify the speech channel, we feed this channel with the extracted Mel-Cepstral Coefficients (MFCC) of the auditory information. Evidence [27] shows that the use of MFCC is suitable for speech representation, but does not provide much information when describing music. MFCCs are described as the coefficients derived from the cepstral representation of an audio sequence, which converts the power spectrum of an audio clip into the Mel-scale frequency. The Mel scale showed to be closer to the human auditory system's response than the linear frequency [28]. This channel is composed of three layers, each one with one-dimensional filters. The first has 5 filters, with a dimension of 1x3, the second one has 10 filters with a dimension of 1x3 and the third one 20 filters with a dimension of 1x2. All three layers apply pooling with a receptive field of 1x2.

When trying to describe general music information, spectral representations, such as power spectrograms, showed good results [29]. Power spectrograms are calculated from smaller sequences of audio clips, by applying a discrete Fourier transform in each clip. This operation describes the distribution of frequency components on each clip and we use it as input for our Music channel. The Music channel is composed of two layers, the first one with 10 filters, and each one with a dimension of 5x5. The second layer has 20 filters, with a dimension of 3x3. Both layers implement pooling, with a receptive field of 2x2. The Speech channel is composed of three layers, each one with one-dimensional filters. The first has 5 filters, with a dimension of 1x3, the second one has 10 filters with a dimension of 1x3 and the third one 20 filters with a dimension of 1x2. All three layers apply pooling with a receptive field of 1x2.

The Cross-channel applied to our Auditory stream has one layer, with 30 filters, each one with a dimension of 2x2, without the application of pooling. To be able to use the Cross-channel, both channels must output data with the same dimensions and our results showed that resizing the Music channel output produced better performance. This can be explained by the fact that the Speech channel depends strongly on the non-locality of the features due to the MFCC representation.

B. Emotion Expression Learning

In the original proposition of the CCCNN with a self-organizing layer, a Self-Organizing Map (SOM) was used to learn emotion concepts. SOMs are neural networks trained in an unsupervised fashion to create a topological grid that represents the input stimuli. In a SOM, each neuron is trained to be a prototype of the input stimuli. So, after training, each neuron will have a strong emotion representation and neurons which are neighbors are related to similar expressions. This means that, after training the SOM with emotion expressions, regions of neurons will code similar expressions representing an emotion concept. For example, a cluster of neurons can be related to a "Happy" expression, while another cluster can be related to "Sad" expressions.

The knowledge of the SOM is represented by its topology, which is fixed. That means that although it is possible to create neural clusters, the SOM will be limited on what it can learn. To learn new information, the topology of the SOM has to be changed and it has to be re-trained. To overcome this limitation, we use a Growing-When-Required Network (GWR) [22].

In a GWR, neurons are added only when necessary, and without any pre-defined topology. This allows the model to have a growing mechanism, increasing and decreasing the number of neurons, and their positions, when required. This makes the model able to represent data with an arbitrary number of samples and introduces the capability of dealing with novelty detection. Similar to a SOM, the GWR also uses best-matching units (BMU) to identify which of the model's neuron has the most similar representation to a certain input.

The behavior of the GWR when iterating over a training set shows the emergence of concepts. In the first epochs the network has an exponential grow in the number of neurons, but after achieving a topology that models the data, it mostly converges. This behavior changes when a new set of training samples is presented to the network. If that new set does not match with some particular region of the network, the model adapts to the new data distribution, forgetting and removing old neurons when necessary, and creating new ones. That gives the model a similar behavior to the formation and storage of memory in the brain [30].

The GWR gives our model three important new characteristics: (I) it removes the limitation on the number and topo-

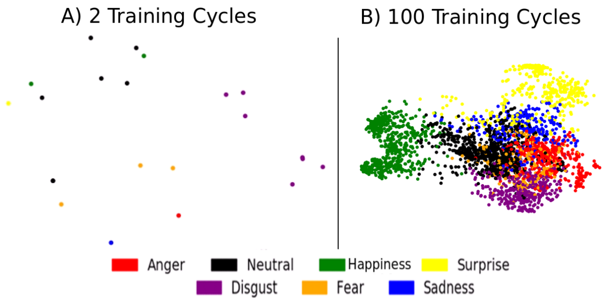

Fig. 2. We proceed to train a Perception GWR, which will maintain our entire representation of multimodal emotion expression perception. The figure illustrates the general network trained with emotion expressions from all our corpora, in the first training cycle on the left, and after 100 ones on the right.

logical structure of the neurons, (II) increases the capability of novelty detection adapting to new expressions the moment they are presented to the network, and lastly, but most important, (III) it has the capability to learn and forget concepts. That means that we can use our GWR to learn how to associate different expression modalities, identify and learn never seen expressions and cluster them into new emotional concepts, and forget concepts which are not important anymore.

We proceed to implement a GWR (Perception GWR) that represents the general knowledge of our perception architecture, composed of the CCCNN, and is able to identify several different types of expressions. We train this Perception GWR with different expression representations coming from the cross-channel layer of the CCCNN in a way that it produces the most general representation as possible. Figure 2 illustrates our Perception GWR in the first interaction of a training routine, on the left, and in the last interaction, on the right. It is possible to see that the network created clusters by itself, as we do not enforce any topological structure.

III. Affective Memory

Although much research on emotion has been done in the past centuries, there is no consensus in the literature on how to define it. The term emotion evolved from the idea of passion around the 16th century [31], and since it can be defined as different concepts such as intense feelings towards another person [32], the current mental state of an entity [33] or even a set of physical and physiological responses to a particularly meaningful event [34]. The latter definition goes further and defines emotion into three different constructs:

- **Feelings** are a subjective representation of emotions which are experienced or perceived by one individual in a certain instant, are usually related to short-term memory.
- **Moods** are affective states, which last longer than feelings, but are less intense.
- **Affect** is a long-term memory relating feelings to objects, persons or events.

These constructs relate directly to the idea of affective memory, which is a concept that defines emotional attributes to differently experienced events by a person [35]. When we perceive a new event, this information is stored in our brain together with an emotional attribute, such as arousal or valence, and recent research shows that events with higher arousal appear to increase the likelihood of being retained in long-term memory [36], [37].

Our model is inspired by the concept of the different emotion constructs [34] and creates different emotion representations based on what is perceived when it was perceived and how it was perceived. Our affective memory model builds on our perception Cross-Channel Convolution Neural Network [17]: First, we use our Perception GWR as a first-step perception mechanism, then we introduce the use of individual GWR networks as different affective memories.

Training the GWR with different expressions gives us a very powerful associative tool which will adapt to the expressions which are presented to it. By adapting the learning and forgetting factors of the GWR we can determine how long the network will keep the learned information, simulating different stages of the human memory process. For example, training a GWR to forget quickly will make it associate and learn local expressions, in a similar way that the encoding stage works. By decreasing the forgetting factor of the network, it is possible to make it learn more expressions, meaning that it can adapt its own neurons topology to a set of expressions that was presented in a mid- to long-time span.

Figure 3 illustrates a GWR architecture used to represent an affective memory for a video sequence. We first proceed to use the Perception GWR to detect which expressions have been performed, and we feed this information to our Affective Memory GWR. We use PCA to create a visualization of the GWR neurons. In the beginning, represented by the topology on the left, it is possible to see that the network memorized most of the neutral concepts. However, at the end, different concepts were represented in reason to the forget factor of our Affective Memory. By changing the forgetting factor of this network, we can let it learn the expressions on the whole video, or just in one part of it.

Using the GWR we can create several kinds of emotional memory of what was perceived. By having other GWRs, with different learning and forgetting factors, we can simulate several types of emotional memory: short- and long-term memory, but also personal affective memory, related to a scene, person or object, and even mood. By feeding each of these memories with the Perception GWR, we can create an end-to-end memory model which will learn and adapt itself based on what was perceived. The Perception GWR can learn new expressions if presented with such, and each of the specific memories will adapt to it in an unsupervised fashion. We then propose our final model containing two affective memories: Affective Memory GWR and Working Memory GWR, reflecting respectively the Feelings and Affect constructs. Figure 4 illustrates our proposed affective memory model.

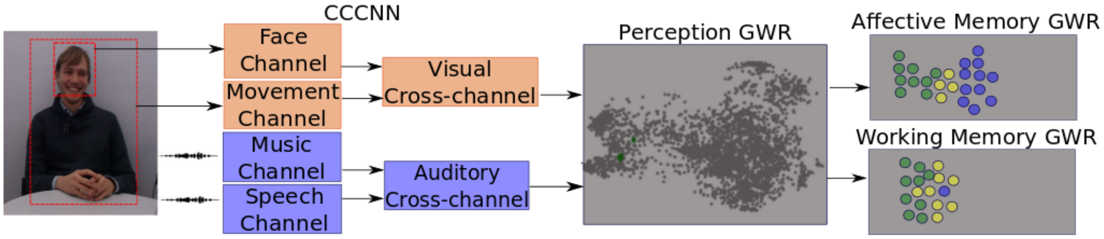

Fig. 4. The final affective memory architecture, which extracts features from the visual and the auditory input with the CCCNN, use the Perception GWR to detect perceived expressions and an Affective Memory GWR and a Working Memory GWR to model long- and short-term memory respectively.

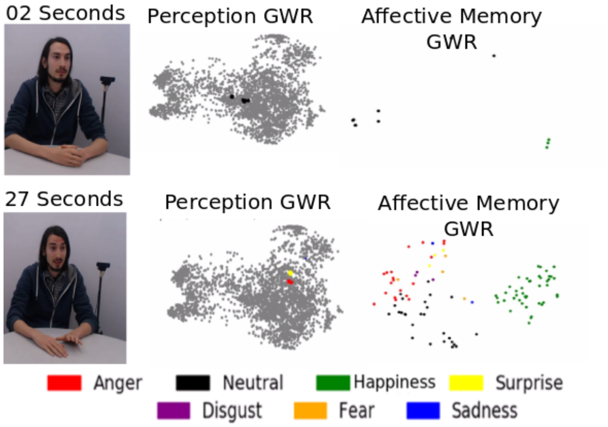

Fig. 3. Using the expressions depicted on the Perception GWR, we proceed to train an Affective Memory GWR for a video. The network on the left illustrates the Affective Memory on the start of the video (02 seconds) and on the right at the end of the video (06 seconds). The colored neurons in the Perception GWR indicate which neurons were activated when the expression is presented and the emotion concept associated with them. The colored neurons on the Affective Memory indicate which emotion concepts these neurons code.

IV. Experimental Setup

To evaluate our model, we make use of two different corpora: The Emotion-Recognition-In-the-Wild-Challenge (EmotiW) and our KT Emotional Interaction Corpus. The EmotiW corpus is part of the Emotion Recognition in the Wild Challenge, and it is considered one of the most challenging datasets on emotion determination. We choose it to evaluate how the performance of our model in such a difficult task is, and how it compares to state-of-the-art methods. The KT Emotional Interaction Corpus was recorded to evaluate the capability of our model to identify natural human behavior over different scenes.

The Emotion-Recognition-In-the-Wild-Challenge (EmotiW) [18] contains video clips extracted from different movies and organized into seven classes: "Anger", "Disgust", "Fear", "Happiness", "Neutral", "Sadness" and "Surprise". A total of 1000 videos with different lengths is available, separated into training and validation sets. The test set is available, but without any label, and includes 700 additional videos. Therefore, we only evaluate our model on the validation set. This challenge is recognized as one of the most difficult tasks for emotion recognition because the movie scenes contain very cluttered environments, occluded faces, speech, music, sound effects, more multiple speakers and even animals. The dataset is separated into pre-defined sets for training, validation and testing the algorithms. We proceed to train our network using the training set and evaluating its performance on the validation set.

The KT Emotional Interaction Corpus contains videos of the interaction between two humans. An improvisation game was recorded, where two subjects are seated in front of each other, across a table. We created two roles: the active and passive subject. To the active subject, we give one instruction and based on this instruction he has to start an improvised dialogue with the passive subject. The instructions were based on five topics:

- Lottery: Tell the other person he or she won the lottery.
- Food: Introduce to the other person a very disgusting food.
- School: Tell the other person that his/her school records are gone.
- Pet: Tell the other person that his/her pet died.
- Family: Tell the other person that a family member of him/her is in the hospital.

These topics were selected in a way to provoke interactions related to at least one of the universal expressions each: "Happiness", "Disgust","Anger", "Fear", and "Sadness". None of the subjects was given any information about the nature of the analyses, to not bias them in their expressions.

Two Logitech HD Pro C920 cameras were placed in a position that they captured the torsos of the persons seated on the chairs. Each camera recorded a video at a resolution of 1024x768 and a framerate of 30FPS. Each participant had a Bluetooth Sennheiser EZX 80 microphone attached to his/her shirt, allowing to record an individual audio channel for each participant. Figure 5 illustrates an example of the recordings.

After each dialogue session, the role of the active subject was given to the previously passive subject and a new topic was assigned. For each pair of participants, a total of five dialogue sessions was recorded, one for each topic, and each one lasting between 30 seconds and 2 minutes. Although the topics were chosen to provoke different expressions, it was the case that in some dialogues none of the previously mentioned emotional concepts was expressed.

The recordings included a total of 7 sessions, with 14 different subjects, two participating in each session. Each

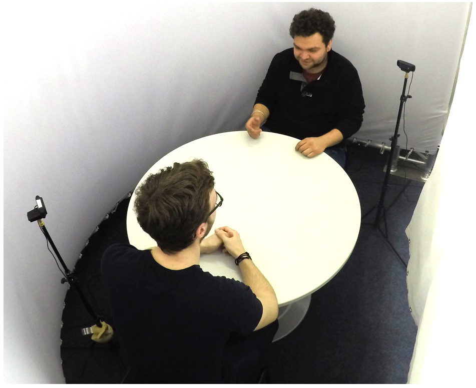

Fig. 5. An example of the recording scenario. One of the participants is chosen as the active subject and one of the five topics is given to him/her.

session had 6 dialogues, one per topic and an extra one where the two subjects introduced each other using a fake name. Each subject only participated in one session, meaning that no subject repeated the experiment. A total of 84 videos was recorded, one for each subject in each dialogue, with a total of 1h05min of recordings. Each video had a different length, with the longest one having 2 minutes and 8 seconds and the shortest one with 30 seconds. To annotate the videos, we cut them into 10s clips and a total of 39 annotators evaluated them using the six universal emotions plus neutral.

A. Experiments

To evaluate our model, we propose two experiments: a performance experiment and a behavior analysis experiment.

1) Performance Experiments: After training the CCCNN with the multimodal expressions, we train our Perception GWR and use the learned representation to classify the expressions. We connect the output of each of the Cross-channels with the GWR and use the hidden and softmax layer to classify the BMU of the perceived emotion.

We follow the same learning protocol which was established in our previous research to train the CCCNN [17]. The GWR parameters were chosen based on the result of an empirical search on the learning and forgetting rates.

We perform this experiment 30 times with both the EmotiW and KT Emotional Interaction Corpus. For each dataset, we calculate the mean accuracy of each category of emotion when using only the CCCNN, the CCCNN+SOM, and the CCCNN+GWR.

2) Behavior Analysis Experiments: To evaluate the behavior using each of our affective memories, we present all the videos of the KT Emotional Interaction Corpus to the model, and let each memory model learn without restriction. Each neuron on the memories will code one expression representation, and we proceed to use the hidden and softmax layers of our CCCNN to create an emotion concept classification. At the end, we have one Affective Memory for each subject and a Working Memory for each topic.

We proceed to calculate the intraclass correlation coefficient between the neurons in each memory and the annotator's opinion for each of the subjects and each of the topics. We then calculate the mean of this correlation as a measure of how far the network memory was from what the annotators perceived.

TABLE I
MEAN ACCURACY, IN PERCENTAGE, FOR THE MULTIMODAL REPRESENTATION IN THE VALIDATION SET OF THE EMOTIW CORPUS. THE RESULTS ARE FOR THE CCCNN AND THE SOM.

Class	CCCNN	CCCNN+SOM	CCCNN+GWR
Anger	80.3	85.3	86.4
Disgust	23.4	30.3	32.6
Fear	30.8	32.1	35.4
Happiness	81.2	82.3	85.2
Neutral	68.7	67.3	67.1
Sadness	24.5	31.7	33.8
Surprise	14.0	17.6	17.5
Mean	46.1	49.5	51.1

TABLE II
REPORTED ACCURACY, IN PERCENTAGE, FOR THE CCCNN TRAINED WITH THE KT EMOTION INTERACTION CORPUS CORPUS WITH A SOM AND WITH A GWR.

Class	CCCNN	CCCNN+SOM	CCCNN+GWR
Anger	80.2	85.4	88.6
Disgust	87.3	91.3	91.0
Fear	71.5	79.0	80.7
Happiness	83.8	92.3	93.2
Neutral	72.8	80.5	89.7
Surprise	81.9	86.7	88.6
Sadness	82.7	87.1	93.2
Mean	80.0	86.0	89.3

V. RESULTS

A. Performance Experiments

The experiments with the EmotiW corpus can be found in Table I. It is possible to see that the GWR could perceive more complex information better, such as "Sadness" and "Fear" clips, which are usually more difficult to recognize by humans. In overall, the GWR showed a better adaptability in most of the categories, improving the general accuracy of the model.

Evaluating the CCCNN with a SOM and with a GWR produced the results shown in Table II. It is possible to see that the GWR increased the accuracy of expressions such as "Fear", "Happiness", and "Sadness" more than the others. Mostly this happens because these expressions display a high degree of variety in the dataset, which is easily perceived by the GWR, by adapting its topology to the new expressions.

B. Behavior Analysis Experiments

The interclass correlation coefficients per topic, represented by the Working Memory GWR, are presented in Table III. It is possible to see high correlations for at least two scenarios: Lottery and Food. These two scenarios were the ones with a stronger correlation also within the annotators, and possibly the ones where the expressions were most easily distinguishable for all the subjects.

The correlation coefficients calculated on the Affective Memory GWR are shown in Table IV. Here, it is possible

TABLE III
INTERCLASS CORRELATION COEFFICIENT OF OUR WORKING MEMORY GWR TRAINED WITH THE KT EMOTIONAL INTERACTION CORPUS

Lottery	Food	School	Family	Pet
0.84	0.71	0.47	0.52	0.53

TABLE IV
INTERCLASS CORRELATION COEFFICIENT OF THE AFFECTIVE MEMORY GWR TRAINED WITH THE KT EMOTIONAL INTERACTION CORPUS

Session	2		3		4		5	
Subject	S0	S1	S0	S1	S0	S1	S0	S1
-	0.79	0.67	0.74	0.79	0.61	0.74	0.67	0.59
Session	6		7		8			
Subject	S0	S1	S0	S1	S0	S1		
-	0.68	0.87	0.68	0.63	0.64	0.76		

to see that for most of the subjects the network presented a positive correlation, while only a few presented a very good one. Also, it is possible to see that the correlations obtained by the emotion concept were again the highest.

VI. DISCUSSION

We can relate the behavior of our model to two memory constructs: **Feelings** and **Affect**. Our Perception GWR relates directly to the **Feelings** construct, as it models the subjective representation of perceived emotions in a certain instant. Our Affective and Working Memories relate directly to the **Affect** construct, which defines the feelings of our model towards a specific person or event in a long-term relation.

In the following sections we discuss in detail our Perception GWR and our different memories.

A. Perception GWR

The EmotiW dataset is still one of the most challenging datasets on emotion recognition tasks. Using our GWR model as a perception mechanism improved the accuracy of the CCCNN model, showing that the GWR actually adapted better to the data than the SOM. When compared to state-of-the-art results on the same data, our GWR method stands as one of the best results, as shown in Table V.

B. Affective Memory

Using Growing When Required Networks (GWR), it was possible to introduce a novel affective memory to our network, which could be adjusted to learning and forgetting in different time steps. This allowed us to create individual memories which are related to particular persons, situations and periods of time, simulating different affective memory mechanisms.

TABLE V
PERFORMANCE OF STATE-OF-THE-ART APPROACHES ON THE EMOTIW DATASET. ALL THE RESULTS CALCULATE THE MEAN ACCURACY ON THE VALIDATION SPLIT OF THE DATASET.

Methodology	Accuracy
[10]	48.53
[11]	41.1
[18]	28.19
[38]	53.80
CCCNN	46.1
CCCNN+SOM	49.5
CCCNN+GWR	51.1

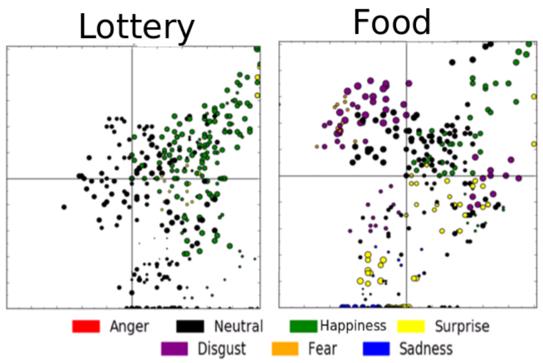

Fig. 6. This plot shows two Working Memory GWRs after trained with all the videos of the KT Emotional Interaction Corpus. We proceed to create one memory for each scenario.

The affective memory mechanisms presented in this paper can be used as a dynamic self-learning strategy for emotion perception and representation. As a clear example, a robot can have a different affective memory of a person which interacts with it on a daily basis, when compared to someone that it just met. Another advantage of using this model is the capability to capture human behavior over different perceived experiences.

In Figure 6, it is exemplified how the Working Memory GWR for two of the topics of the KT Emotional Interaction Corpus: food and lottery. It is possible to see also how some emotional concepts are present in each scenario. While in the food scenario, many “Disgust” annotations are present, in the lottery scenario the interactions are labeled mostly as “Happiness” or “Surprise”.

VII. CONCLUSION

Different approaches in the affective computing area deal with emotion recognition and description. Most of them fail when categorizing spontaneous and natural expressions due to the lack of adaptability or to the capability to generalize to different persons and scenarios. In this paper, we make use of the power of the Cross-channel Convolution Neural Network (CCCNN) to describe multimodal and spontaneous emotion expressions together with the adaptability of the Growing-When-Required network (GWR) to learn and forget information in an unsupervised fashion.

Our model builds on a previously proposed model, removing the restrictions of the limited topology of the Self-Organizing Maps (SOM), but also introducing the use of different GWRs as affective memory mechanisms. These memories can detect and learn emotion behavior from different persons and situations at the same time, and keep a dynamic behavior while adapting to new information.

We evaluated our model in two tasks: performance in emotion recognition and behavior analysis, and demonstrated that our model is competitive with state-of-the-art methods.

We also introduce the novel KT Emotion Interaction Corpus, which contains interactions between two humans and emotion annotations.

Currently, the GWR neurons provide an associative relation between the visual and auditory modalities of the CCCNN but cannot deal with information conflict between the modalities, which would be an interesting aspect of the network to be developed in future research. Adding such mechanism would increase the robustness and the capability of the model to learn multimodal expressions. Also creating a mechanism on which the memories can modulate their functioning to lead to a mood memory which could improve the adaptability of the model to self-perception mechanisms.

ACKNOWLEDGEMENTS

This work was partially supported by German Research Foundation DFG under project CML (TRR 169).

REFERENCES

[1] F. Foroni and G. R. Semin, "Language that puts you in touch with your bodily feelings the multimodal responsiveness of affective expressions," *Psychological Science*, vol. 20, no. 8, pp. 974–980, 2009.

[2] P. Rani and N. Sarkar, "Emotion-sensitive robots - a new paradigm for human-robot interaction," in *Humanoid Robots, 2004 4th IEEE/RAS International Conference on*, vol. 1, Nov 2004, pp. 149–167 Vol. 1.

[3] D. Bandyopadhyay, V. C. Pammi, and N. Srinivasan, "Chapter 3 - role of affect in decision making," in *Decision Making Neural and Behavioural Approaches*, ser. Progress in Brain Research, V. C. Pammi and N. Srinivasan, Eds. Elsevier, 2013, vol. 202, pp. 37 – 53.

[4] R. W. Picard and R. Picard, *Affective computing*. MIT Press Cambridge, 1997, vol. 252.

[5] L. C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *Information, Communications and Signal Processing, 1997. ICICS., Proceedings of 1997 International Conference on*, vol. 1. IEEE, 1997, pp. 397–401.

[6] I. A. Essa and A. P. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 757–763, 1997.

[7] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[8] S. Afzal and P. Robinson, "Natural affect data - collection and annotation in a learning context," in *3rd International Conference on Affective Computing and Intelligent Interaction.*, September 2009, pp. 1–7.

[9] F. Ringeval, S. Amiriparian, F. Eyben, K. Scherer, and B. Schuller, "Emotion recognition in the wild: Incorporating voice and lip activity in multimodal decision-level fusion," in *Proceedings of the 16th International Conference on Multimodal Interaction*, ser. ICMI '14. New York, NY, USA: ACM, 2014, pp. 473–480.

[10] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 494–501.

[11] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, c. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, M. Côté, K. R. Konda, and Z. Wu, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ser. ICMI '13. New York, NY, USA: ACM, 2013, pp. 543–550.

[12] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 503–510.

[13] B. Sun, L. Li, G. Zhou, and J. He, "Facial expression recognition in the wild based on multimodal texture features," *Journal of Electronic Imaging*, vol. 25, no. 6, pp. 407–427, 2016.

[14] J. A. Russell, "Core affect and the psychological construction of emotion." *Psychological review*, vol. 110, no. 1, p. 145, 2003.

[15] L. Cahill and J. L. McGaugh, "A novel demonstration of enhanced memory associated with emotional arousal," *Consciousness and Cognition*, vol. 4, no. 4, pp. 410–421, 1995.

[16] M. W. Eysenck, "Arousal, learning, and memory." *Psychological Bulletin*, vol. 83, no. 3, p. 389, 1976.

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

[18] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 461–466.

[19] J. K. Hamlin, "Moral judgment and action in preverbal infants and toddlers evidence for an innate moral core," *Current Directions in Psychological Science*, vol. 22, no. 3, pp. 186–193, 2013.

[20] F. Pons, P. L. Harris, and M. de Rosnay, "Emotion comprehension between 3 and 11 years: Developmental periods and hierarchical organization," *European journal of developmental psychology*, vol. 1, no. 2, pp. 127–152, 2004.

[21] M. Lewis, *Children's emotions and moods: Developmental theory and measurement*. Springer Science & Business Media, 2012.

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

[23] M. E. Kret, K. Roelofs, J. Stekelenburg, and B. de Gelder, "Emotional signals from faces, bodies and scenes influence observers' face expressions, fixations and pupil-size," *Frontiers in human neuroscience*, vol. 7, p. 810, 2013.

[24] D. C. V. Essen and J. L. Gallant, "Neural mechanisms of form and motion processing in the primate visual system," *Neuron*, vol. 13, no. 1, pp. 1 – 10, 1994.

[25] P. Barros, G. Parisi, D. Jirak, and S. Wermter, "Real-time gesture recognition using a humanoid robot with a deep neural architecture," in *Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on*, Nov 2014, pp. 646–651.

[26] G. Hickok, "The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model," *Journal of communication disorders*, vol. 45, no. 6, pp. 393–402, 2012.

[27] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A.-R. Mohamed, G. Dahl, and B. Ramabhadran, "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, vol. 64, pp. 39 – 48, 2015.

[28] M. Pitz and H. Ney, "Vocal tract normalization as linear transformation of MFCC." in *INTERSPEECH*, vol. 3, 2003, pp. 1445–1448.

[29] J. George and L. Shamir, "Unsupervised analysis of similarities between musicians and musical genres using spectrograms," *Artificial Intelligence Research*, vol. 4, no. 2, p. 61, 2015.

[30] Y. Dudai, "The neurobiology of consolidations, or, how stable is the engram?" *Annu. Rev. Psychol.*, vol. 55, pp. 51–86, 2004.

[31] T. Dixon, *From passions to emotions: The creation of a secular psychological category*. Cambridge University Press, 2003.

[32] D. Hume, "Emotions and moods," *Organizational behavior*, pp. 258–297, 2012.

[33] B. Fehr and J. A. Russell, "Concept of emotion viewed from a prototype perspective." *Journal of experimental psychology: General*, vol. 113, no. 3, p. 464, 1984.

[34] E. Fox, *Emotion Science: An Integration of Cognitive and Neuroscience Approaches*. Palgrave Macmillan, 2008.

[35] K. S. LaBar and E. A. Phelps, "Arousal-mediated memory consolidation: Role of the medial temporal lobe in humans," *Psychological Science*, vol. 9, no. 6, pp. 490–493, 1998.

[36] E. A. Kensinger, "Remembering emotional experiences: The contribution of valence and arousal," *Reviews in the Neurosciences*, vol. 15, no. 4, pp. 241–252, 2004.

[37] E. A. Phelps, S. Ling, and M. Carrasco, "Emotion facilitates perception and potentiates the perceptual benefits of attention." *Psychological science*, vol. 17, no. 4, pp. 292–299, Apr. 2006.

[38] A. Yao, J. Shao, N. Ma, and Y. Chen, "Capturing au-aware facial features and their latent relations for emotion recognition in the wild," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 451–458.