Special issue on Grounding Emotions in Robots

Adaptive Behavior

Developing crossmodal expression recognition based on a deep neural model

Adaptive Behavior
2016, Vol. 24(5) 373–396

DOI: 10.1177/1059712316664017
adb.sagepub.com
SAGE

Pablo Barros and Stefan Wermter

Abstract

A robot capable of understanding emotion expressions can increase its own capability of solving problems by using emotion expressions as part of its own decision-making, in a similar way to humans. Evidence shows that the perception of human interaction starts with an innate perception mechanism, where the interaction between different entities is perceived and categorized into two very clear directions: positive or negative. While the person is developing during childhood, the perception evolves and is shaped based on the observation of human interaction, creating the capability to learn different categories of expressions. In the context of human–robot interaction, we propose a model that simulates the innate perception of audio–visual emotion expressions with deep neural networks, that learns new expressions by categorizing them into emotional clusters with a self-organizing layer. The proposed model is evaluated with three different corpora: The Surrey Audio–Visual Expressed Emotion (SAVEE) database, the visual Bi-modal Face and Body benchmark (FABO) database, and the multimodal corpus of the Emotion Recognition in the Wild (EmotiW) challenge. We use these corpora to evaluate the performance of the model to recognize emotional expressions, and compare it to state-of-the-art research.

Keywords

Crossmodal learning, emotion expression recognition, convolution neural network, self-organizing maps

1 Introduction

The communication task is one of the most important tasks within the area of Human–Robot Interaction (HRI). The most necessary skills of human–human communication are the capability to perceive, understand and respond to social interactions, usually determined through affective expressions, as discussed by Foroni and Semin (2009). As discussed by Cabanac (2002), there is no consensus in the literature to define emotional expressions. However, Ekman and Friesen (1971) developed a study that shows that emotion expressions have a universal understanding, independent of gender, age and cultural background. They established the six universal emotions: disgust, fear, happiness, surprise, sadness and anger. Although they show that these emotions are commonly inferred from expressions by most people, the concept of spontaneous expressions increases the complexity of the expression representation. Humans usually express themselves differently, sometimes even combining one or more characteristics of the so-called *universal emotions*. Several researchers have built their own categories of complex emotional states, with concepts such as confusion, surprise, and concentration as exhibited by Afzal and Robinson (2009). To define spontaneous emotions, the observation of several characteristics, and among them, facial expressions, movement and auditory signals, has been shown to be necessary, as demonstrated by Kret, Roelofs, Stekelenburg, and de Gelder (2013). They perform a psychological analysis on the observation of the whole body for emotional expressions. They show that face expression alone may contain misleading information, especially when applied to interaction and social scenarios.

The observation of different modalities, such as body posture, motion, and speech intonation, improved the determination of the emotional state of the subjects. This was demonstrated in the computational system of Castellano, Kessous, and Caridakis (2008), where they process facial expression, body posture, and speech,

Department of Informatics, University of Hamburg, Germany

Corresponding author:
Pablo Barros, Department of Informatics, University of Hamburg, Knowledge Technology, Vogt-Kölln-Straße 30, D-22527 Hamburg, Germany.
Email: barros@informatik.uni-hamburg.de

extracting a series of features from each modality and combining them into one feature vector. Although they show that when different modalities are processed together they present a better recognition accuracy, the individual extraction of each modality does not model the correlation between them, which could be found when processing the modalities together as one stream. The same principle was also found for visual-only modalities, in the works of Gunes and Piccardi (2009) and Chen, Tian, Liu, and Metaxas (2013) and audio-only modalities, in the works of Ringeval, Amiriparian, Eyben, Scherer, and Schuller (2014), Jin, Li, Chen, and Wu (2015) and Liu, Chen, Li, and Zhang (2015). However, all these works deal with a set of restricted expression categorizations, which means that if a new emotion expression is presented to these systems, they must be re-trained and a new evaluation and validation of the whole system need to be done.

Dealing with a set of restricted emotions is a serious constraint to HRI systems. Humans have the capability to learn emotion expressions and adapt their internal representation to a newly perceived emotion. This is explained by Hamlin (2013) as a developmental learning process. Her work shows that human babies perceive interactions into 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. The same process was also described by Harter and Buddin (1987), Lewis (2012) and Pons, Harris, and de Rosnay (2004).

The most successful way to represent data is the one done by the human brain, as explained by Adolphs (2002). He discusses how the human brain recognizes emotional expressions from visual and auditory stimuli, correlating information from different areas. The brain correlates past experiences, movements and face expressions with perceived sounds and voices. The brain is also capable of integrating this multimodal information and generates a unique representation of the visual and auditory stimuli. The simulation of this process in computer systems can be achieved by neural models, particularly ones that are able to create a hierarchy of feature representations such as Convolutional Neural Networks (CNNs), introduced by Lecun, Bottou, Bengio, and Haffner (1998) and used for different visual tasks, as demonstrated by the works of Lawrence, Giles, Tsoi, and Back (1997), Karnowski, Arel, and Rose (2010) and Khalil-Hani and Sung (2014), and auditory tasks in the works of Sainath et al. (2015), Li, Chan, and Chun (2010) and Schluter and Bock (2014).

This paper proposes an automatic emotion recognition system that is inspired by the learning aspects of human emotion expression perception. The first step is to create a perception representation for different modalities that preserves the information of each individual modality, but also models the correlations within them. From there, a computational model for developmental emotional perception gives the system the capability to adapt its own perception mechanisms to different people and expressions.

In this paper, we propose the use of a Crosschannel Convolutional Neural Network (CCCNN) by Barros, Weber, and Wermter (2015b) to integrate auditory and visual stimuli into one representation. This network represents the information in separated channels and applies a crosschannel to integrate different input modalities without losing the unique representation from each stimulus. The second step is accomplished by the application of a self-organizing layer on top of the CCCNN in order to establish separation boundaries to the perceived expressions.

To evaluate our model, two different sets of experiments are performed. The first one relates to the emotion expression representation and the second one to the emotion expression learning. In the first set of experiments, we train and evaluate our CCCNN for multimodal emotion expression recognition. In the second step, we use our self-organizing layer to learn new expressions. We use a total of three different corpora: The Bi-modal Face and Body benchmark database (FABO) for visual emotion expression, the Surrey Audio–Visual Expressed Emotion (SAVEE) database for auditory expressions, and the Emotion-Recognition-In-the-Wild-Challenge (EmotiW) for multimodal expressions.

This paper is organized as follows: The next section shows the proposed model for emotion expression representation, describing how it deals with the two stimuli modalities. Section 2 extends our model and adapts it to the developmental expression learning strategy. Section 3 describes our experimental methodology, shows our results and compares them with state-of-the-art solutions. Discussion of the results, the role of each modality, the learning strategy, and a discussion about human emotion representations in the proposed model are given in Section 4. Conclusions and future works are presented in the last section.

2 Emotion expression representation

Our model deals with multimodal stimuli, and takes into consideration visual, primary face expressions and body movements, and auditory information. It is implemented as a CCCNN and it extracts hierarchical features from the two modalities. The complex representation varies depending on the presented stimuli, and each hierarchical layer of the network learns a different level of abstraction. That means that the deeper layers will have a full representation of the input while

the first layers will have a local representation of some regions or parts of the input stimuli.

2.1 Convolutional neural network

A CNN is composed of several layers of convolution and pooling operations stacked together. These two operations simulate the responses of simple and complex cell layers discovered in visual area V1 by Hubel and Wiesel (1959). In a CNN, the abstraction of the simple cells is represented by the use of convolution operations, which use local filters to compute high-level features from the input stimulus. The pooling operation abstracts the complex cells by increasing the spatial invariance of the stimulus by pooling simple cell units of the same receptive field in previous layers.

Every layer of a CNN applies different filters, which increases the capability of the simple cells to extract features. Each filter is trained to extract a different representation of the same receptive field, which generates different outputs, or feature maps, for each layer. The complex cells, pool units of receptive fields in each feature map. These feature maps are passed to another layer of the network, and because of the complex cells' pooling mechanism, each layer applies a filter in a receptive field, which contains the representation of a larger region of the initial stimulus. This means that the first layer will output feature maps that contain representations of one region of the initial stimulus, and deeper layers will represent larger regions. At the end, the output feature map will contain the representation of the whole stimulus.

Each set of filters in the simple cell layers acts in a receptive field in the input stimulus. The activation of each unit $u_{n,c}^{x,y}$ at (x,y) position of the nth feature map in the cth layer is given by

$$u_{n,c}^{x,y} = \max(b_{nc} + S, 0) \tag{1}$$

where $\max(\cdot, 0)$ represents the rectified linear function, which was shown to be more suitable for training deep neural architectures, as discussed by Glorot, Bordes, and Bengio (2011). b_{nc} is the bias for the nth feature map of the cth layer and S is defined as

$$S = \sum_{m=1}^{M} \sum_{h=1}^{H} \sum_{w=1}^{W} w_{(c-1)m}^{hw} u_{(c-1)m}^{(x+h)(y+w)} \tag{2}$$

where m indexes over the set of filters M in the current layer, c, which is connected to the input stimulus on the previous layer $(c-1)$. The weight of the connection between the unit $u_{n,c}^{x,y}$ and the receptive field with height H and width W of the previous layer $c-1$ is $w_{(c-1)m}^{hw}$.

A complex cell is connected to a receptive field in the previous simple cell, reducing the dimension of the feature maps. Each complex cell outputs the maximum activation of the receptive field $u(x,y)$ and is defined as

$$a_j = \max_{n \times n}(u_{n,c}(x,y)) \tag{3}$$

where $u_{n,c}$ is the output of the simple cell. In this function, a complex cell computes the maximum activation among the receptive field (x,y). The maximum operation down-samples the feature map, maintaining the simple cell structure.

2.2 Cubic receptive fields

In a CNN, each filter is applied to a single instance of the stimulus and extracts features of a determined region. We can see an emotion expression as a series of single instances of stimuli stacked together. As emotion expressions usually do not contain a strong context dependency, because of the short time in which each emotion is expressed, we use the concept of filtering similar patterns in a stack of stimuli. To do so, we adapt the CNN to apply similar filters on the same region of different images. This concept is obtained by the application of the cubic receptive fields, described by Ji, Xu, Yang, and Yu (2013). In a cubic receptive field, the value of each unit $u_{n,c}^{x,y,z}$ at the nth filter map in the cth layer is defined as

$$u_{n,c}^{x,y,z} = \max(b_{nc} + S_3, 0) \tag{4}$$

where $\max(\cdot, 0)$ represents the rectified linear function, b_{cn} is the bias for the nth filter map of the cth layer, and S_3 is defined as

$$S_3 = \sum_{m} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{r=1}^{R} w_{(c-1)m}^{hwr} u_{(m-1)}^{(x+h)(y+w)(z+r)} \tag{5}$$

where m indexes over the set of feature maps in the $(c-1)$ layer connected to the current layer c. The weight of the connection between the unit $u_{n,c}^{(x,y,z)}$ and a receptive field connected to the previous layer $(c-1)$ and the filter map m is $w_{(c-1)m}^{hwr}$. H and W are the height and width of the receptive field and z indexes each stimulus; R is the number of stimuli stacked together representing the new dimension of the receptive field.

2.3 Shunting inhibition

To learn general features, several layers of simple and complex cells are necessary, which leads to a large number of parameters to be trained. This, put together with the usual necessity of a huge amount of data so that the filters learn general representations, is a big problem shared among deep neural architectures. To reduce the necessity of a deeper network, we introduce the use of shunting inhibitory fields, described by Fregnac, Monier, Chavane, Baudot, and Graham (2003), which improves the efficiency of the filters in learning complex patterns.

Shunting inhibitory neurons are neural-physiological plausible mechanisms that are present in several visual and cognitive functions, as shown by Grossberg (1992). When applied to complex cells, shunting neurons can result in filters that are more robust to geometric distortions, meaning that the filters learn more high-level features. Each shunting neuron S_{nc}^{xy} at the position (x,y) of the nth receptive field in the cth layer is activated as

$$S_{nc}^{xy} = \frac{u_{nc}^{xy}}{a_{nc} + I_{nc}^{xy}} \tag{6}$$

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. The weights of each inhibitory neuron are trained with backpropagation. A passive decay term, a_{nc}, is a defined parameter and it is the same for the whole shunting inhibitory field.

The idea behind the shunting neurons is that they will specify the filters of a layer. This creates a problem when applied to filters that extract low-level features, such as edges and contours. When applied to such filters, the shunting neurons specify these filters, causing a loss on the generalization aspects of the low-level features. However, when applied to deeper layers, the shunting neurons can enhance the capability of the filters to extract strong high-level representations, which could only be achieved by the use of a deeper network.

2.4 Crosschannel learning

To be able to deal with multimodal data, our network uses the concept of the CCCNN by Barros, Jirak, Weber, and Wermter (2015a). In the CCCNN architecture, several channels, each one of them composed of an independent sequence of convolution and pooling layers, are fully connected at the end to a crosschannel layer, which is composed of convolution and pooling layers, and trained as one single architecture. Our architecture is composed of two main streams: a Visual and an Auditory stream.

Our model applies topological convolution, and thus the size of the receptive field has an important impact in the learning process. The receptive fields in our crosschannel need to be large enough to be able to capture the whole concept of the stimulus, and not only part of it. With a small receptive field, our crosslearning will not be able to capture the high-level features.

We apply our crosschannel learning in two streams. Goodale and Milner (1992) describe how the visual cortex is separated into two streams, and how they are integrated in the V4 area. In their model, the ventral and dorsal streams extract different information from the input data, but are used as input to the V4 area. Hickok (2012) describes a similar process occurring in the auditory pathway, where different information is processed by the ventral and dorsal stream, and integrated in the V4 area. Although we are not modeling exactly the same pathway and information as the ones present in the brain, the architecture of our model was developed in a way that resembles the brain's organizational structure. Also, we specify our model to deal with emotion expressions, and not general visual and auditory recognition.

2.5 Visual representation

Inspired by the primate visual cortex model described by Essen and Gallant (1994), our Visual stream has two channels. The first channel is responsible for learning and extracting information about face expressions, which comprises the contour, shape and texture of a face, and mimics the encoding of information in the ventral area of primate visual cortex. The second channel codes information about the orientation, direction and speed of changes within the torso of a person in a sequence of images, similar to the information coded by the dorsal area.

To feed our Visual stream, we must first find the faces on the images. To do so, we use the Viola–Jones face-detection algorithm, proposed in the work of Viola and Jones (2004), which uses an Adaboost-based detection. After finding the face, we create a bounding box to describe the torso movement. We extract face and torso from a sequence of frames corresponding to 1 s and feed them to the network.

To define the input of the Movement channel, we use a motion representation. Feeding this stream with this representation, and not the whole image, allows us to specialize the channel into learning motion descriptors. This way, we can train the network with a smaller amount of data, and use a shallow network to obtain high-level descriptors. Figure 1 displays a common input of our Visual stream, containing examples of the Face and Movement channels.

The Face channel is composed of two convolution and pooling layers. The first convolution layer implements 5 filters with cubic receptive fields, each one with a dimension of $5 \times 5 \times 3$. The second layer implements 5 filter maps, also with a dimension of 5×5, and a shunting inhibitory field. Both layers implement max-pooling operators with a receptive field of 2×2.

The Movement channel implements three convolution and pooling layers. The first convolution layer implements 5 filters with cubic receptive fields, each one with a dimension of $5 \times 5 \times 3$. The second and third channels implement 5 filters, each one with a dimension of 5×5 and all channels implement max-pooling with a receptive field of 2×2. We feed this channel with 1 s of expressions, meaning that we feed the network with 30 frames. We compute the motion representation of every 10 frames, meaning that we feed the Movement channel with 3 motion representations. All the images are resized to 128×96 pixels.

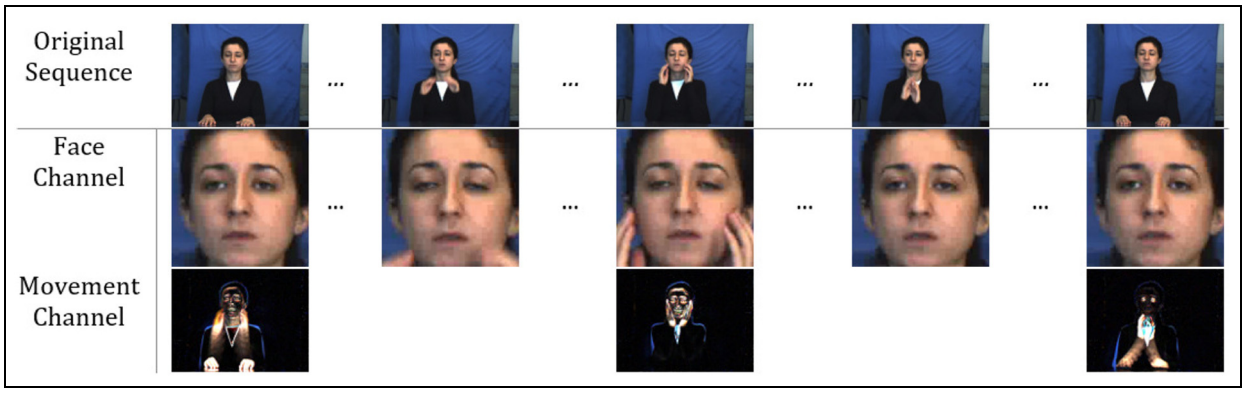

Figure 1. Example of input for the Visual stream. We feed the network with 1 s of expressions, which are processed into 3 movement frames and 9 facial expressions.

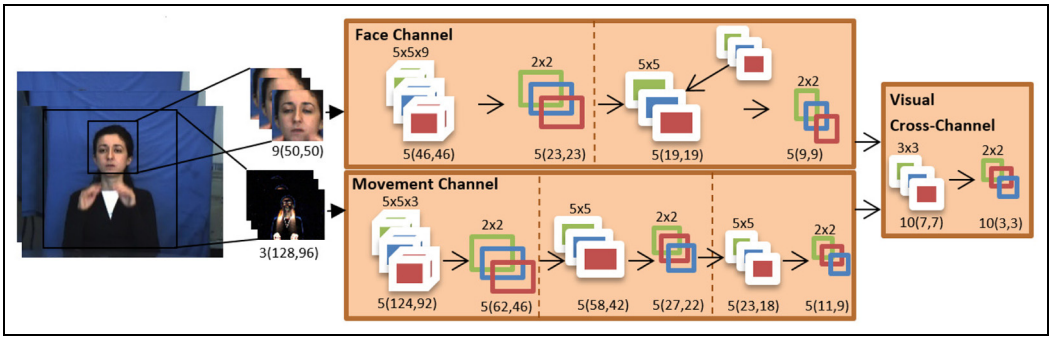

Figure 2. The Visual stream of our network is composed of two channels: the Face and the Movement channels. The Face channel is implemented by two layers, each one with convolution and pooling, and applies inhibitory fields in the second layer, while the Movement channel is implemented by three layers, with pooling and convolution. Both channels implement cubic receptive fields in the first layer. The final output of each channel is fed to a crosschannel which implements convolution and pooling and produces a final visual representation.

We apply a crosschannel to the Visual stream. This crosschannel receives as input the Face and Movement channels, and it is composed of one convolution channel with 10 filters, each one with a dimension of 3×3, and one max-pooling with a receptive field of 2×2. We have to ensure that the input of the crosschannel has the same dimension, to do so we re-size the output representation of the Movement channel to 9×9, the same as the Face channel. Figure 2 illustrates the Visual stream of the network.

2.6 Auditory representation

Hickok (2012) states that the dorsal and ventral pathways in the brain process different auditory information. While the ventral stream deals with speech information, the dorsal one maps auditory sensory representation. In earlier stages of the dorsal stream, the auditory information is decomposed into a series of representations, which are not connected to phonetic representations. We use this concept to separate the perception of auditory information in our network into two channels. One deals mostly with speech signals, and the other with general sounds including music.

Evidence in the work of Sainath et al. (2015) shows that the use of Mel Frequency Cepstral Coefficients (MFCC) is suited for speech representation, but does not provide much information when describing music. MFCCs are described as the coefficients derived from cepstral representation of an audio sequence, which converts the power spectrum of an audio clip into the mel scale frequency. The mel scale has been shown to be closer to human auditory system's response than the linear frequency.

When trying to describe general music information, spectral representations, such as power spectrograms, showed good results, as described in the work of George and Shamir (2015). Power spectrograms are calculated in 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.

To use the auditory representation in CNNs, the MFCCs and power spectrograms are represented as

images. But there is a fundamental difference when dealing with these representations. Usually, the input of CNNs is processed by a filter matrix, which is applied in both, height and width axes. The filter is trained to capture local information of the region where it is applied. When this concept is applied to auditory representation, the idea of learning from a 2D region can generate a problem. In auditory input, each axis represents different information, where usually the x-axis represents time and the y-axis the spectral representation. For the power spectrogram representations, the use of 2D filters was shown to be ideal, because each filter captures the spectral representation in a certain region of the audio clip, as discussed by Hau and Chen (2011).

On the MFCC representation, the use of 2D filters does not work. To Extract the MFCCs, a cosine transformation is applied and this projects each value of the y-axis into the mel frequency space, which may not maintain locality. Because of the topological nature of 2D filters, the network will try to learn patterns in adjacent regions, which are not represented adjacently in the mel frequency domain. Abdel-Hamid et al. (2014) propose the use of 1D filters to solve this problem. The convolution process is the same, but the network applies 1D filters on each value of the y-axis of the image. That means that the filters will learn how to correlate the representation per axis and not within neighbors. Pooling is also applied in 1D, always keeping the same topological structure.

We build our auditory stream based on the speech and music representation. We use two channels, which are connected to a crosschannel. We use audio clips with 1 s as input, and each clip is re-sampled to 16,000 Hz. We compute the power spectrum and the MFCC of the audio clip and feed them to two channels. The power spectrogram is the input of the Music channel, and it is computed over a window of 25 ms with a slide of 10 ms. The frequency resolution is 2048. This generates a spectrogram with 1024 bins, each one with 136 descriptors. We re-size the spectrogram by a factor of 8, resulting in an input size of 128×7. The MFCC is used as input for the Speech channel, and it is calculated over the same window and slide as the power spectrogram. We change the frequency resolution to 1024, which generated a representation with 35 bins each one with 26 descriptors.

The Music channel is composed of two layers, the first one with 10 filters, and each one with a dimension of 5×5. The second layer has 20 filters, with a dimension of 3×3. Both layers implement pooling, with a receptive field of 2×2. The Speech channel is composed of three layers, each one with 1D filters. The first has 5 filters, with a dimension of 1×3, the second one has 10 filters with a dimension of 1×3 and the third one 20 filters with a dimension of 1×2. All three layers apply pooling with a receptive field of 1×2.

The crosschannel applied to our Auditory stream has one layer, with 30 filters, each one with a dimension of 2×2, without the application of pooling. To be able to use the crosschannel, both channels must output data with the same dimensions and our results showed that re-sizing 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. Figure 3 illustrates our Auditory stream.

2.7 Crossmodal representation

To deal with crossmodal learning, we integrate both streams into one Multichannel Convolutional Neural Network architecture. We connect each crosschannel 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-level representation preserved. Figure 4 illustrates our final architecture.

Erhan et al. (2010) show evidence that the use of supervised pre-training steps improves the capability of the filters to tune faster in a specific domain. We follow

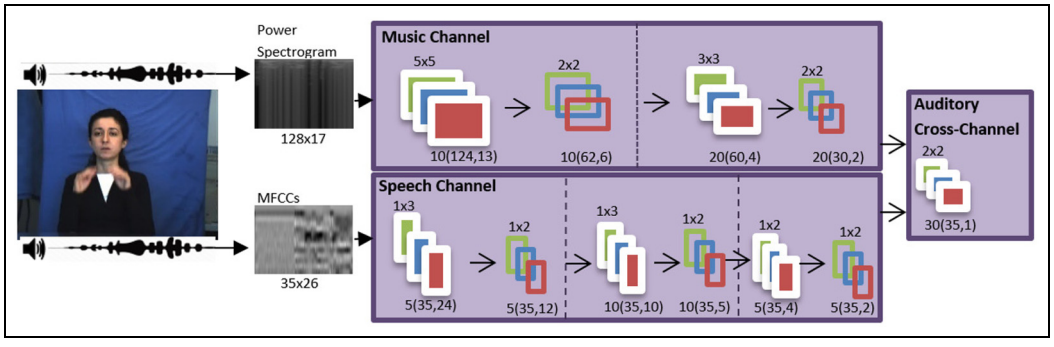

Figure 3. The Auditory stream of our network implements two channels: the Music channel and the Speech channel, which implements filters with one dimension. We feed the network with 1s audio clips, and calculate a power spectrogram as input for the Music channel and MFCCs as input for the Speech channel. The output of both channels is used as input for the auditory crosschannel.

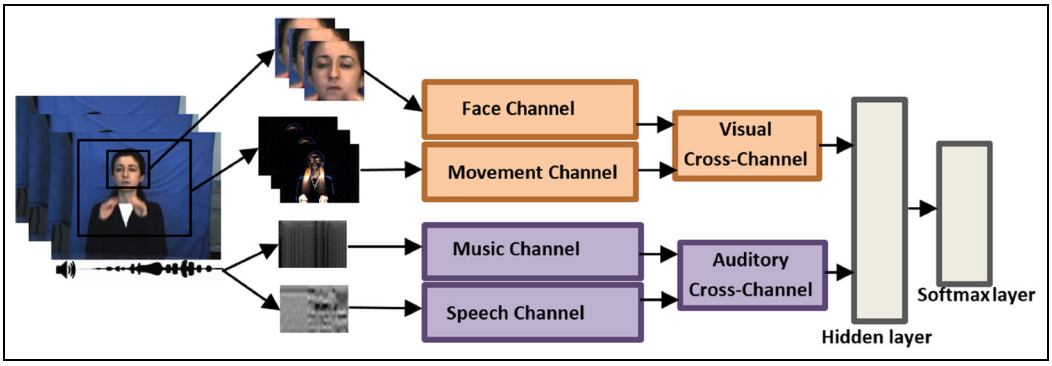

Figure 4. Final crossmodal architecture, which extracts features from visual and auditory input and classifies them in emotion expressions. We connect the outputs of each stream to a fully connected hidden layer and then to a softmax layer, which will give us a classification probability.

this strategy and pre-train each channel of our network to learn specific representation from specific data. After the filters of each channel are trained, we then train our fully connected hidden layer and the softmax layer to classify the input. This strategy allows us to decrease the amount of time needed to train our network and increase the generalization property of our filters.

2.8 Inner representation visualization

CNNs were successfully used in several domains. However, most of the works with CNNs do not explain why the model was so successful. As CNNs are neural networks that learn representation of the input data, the knowledge about what the network learns can help us to understand why these models perform so well in different tasks. The usual method to evaluate the learned representations of neural networks is the observation of the weights matrices, which is not suited for CNNs. Every filter in the convolution layers learns to detect certain patterns in the regions of the input stimulus, and because of the pooling operations, the deeper layers learn patterns that represent a far larger region of the input. That means that the observation of the filters does not give us a reliable way to evaluate the knowledge of the network.

Zeiler and Fergus (2014) proposed the deconvolutional process, which helps to visualize the knowledge of a CNN. In their method, they backpropagate the activation of each neuron to an input, which helps to visualize to which part of the input the neurons of the network are activated for. This way, we can determine regions of neurons that activated for similar patterns, for example, neurons that activate for the mouth and others for the eyes.

In a CNN, each filter tends to learn similar patterns, which indicates that those neurons in the same filter will be activated to similar input structures. Also, each neuron can be activated for very specific patterns, which are not high-level enough for subjective analysis. To improve the quality of our analysis, we apply the concept of creating visualizations for all neurons in one filter by averaging the activation of each neuron in that filter. That allows us to cluster the knowledge of the network in filters, meaning that we can identify if the network has specialized filters, and not specialized neurons. Also, visualizing filters on all layers help us to understand how the network builds the representation, and help us to demonstrate the hierarchical capabilities of CNNs.

The visualizations are a very powerful tool that help us to have an important insight into the knowledge learned by the network. With them, we can validate the parameters of our model, understand what the model learns, and illustrate the advantages of using concepts such as the inhibitory fields and the crosschannels. We also use the visualizations to illustrate which are the most important features, in the network's perspective, for emotion expression, and how they are combined in different modalities. For the auditory channels, the visualizations do not give us an easily understandable indication of what the network learns, different from the visual channels. In the visual channels, we can see which regions of the images the network activates most, but for the auditory channels the input is masked by the transformation of the audio into MFCCs and power spectrograms.

3 Emotion expression learning

To classify emotion expressions is a difficult task: First the observation of various different modalities is necessary. Second, the concept of emotion itself is not precise, and the idea of classifying what another person is expressing based on fuzzy concepts, makes the analysis of such models difficult. Russell (2003) classifies emotion representation into two different model classes: categorical models and dimensional models.

The categorical model separates emotions as discrete concepts, such as the ones proposed by Ekman and

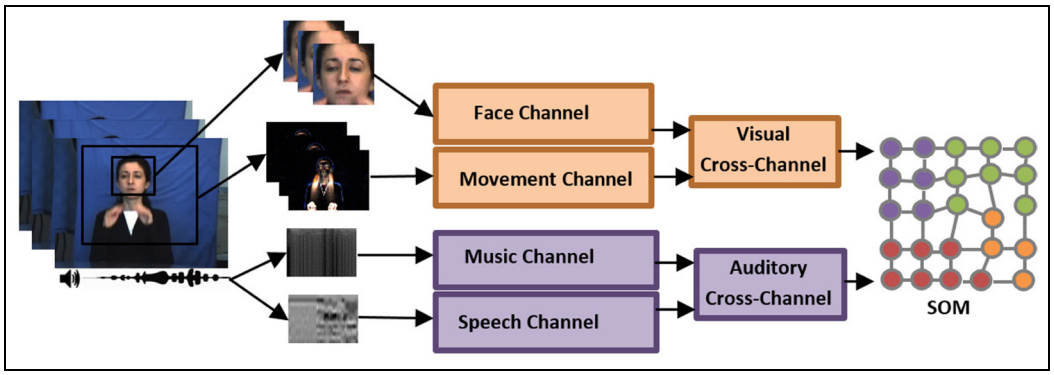

Figure 5. Crossmodal architecture used as input for the SOM. This architecture extracts multimodal features from audio–visual inputs and clusters the representation in different regions, which then represent emotion expressions.

Friesen (1971): disgust, fear, happiness, surprise, sadness and anger. Based on that model, the combination of basic emotions can be expressed as a set of secondary emotions and the idea of tertiary and even ternary emotions was discussed by Sloman (2001). The categorical model became very popular among automatic emotion-recognition systems, because it is easier to train a model where a set of categories is established.

In the dimensional model, the emotion expressions are represented in a two-dimensional space, usually arousal and valence. This dimensional space represents emotions based on their intensity and nature, where a high valence is usually related to positive emotions and a low valence to negative emotions. High arousal is usually related to expressions with high intensity, such as excitement, and low arousal as calm and relaxed expressions. This model has a richer representation of the expressions, without relying on pre-defined categories. Russell (2003) claims that the use of the dimensional model can represent a core affect representation, which models a perpetual affective state that is updated by different stimuli.

A well-known method of describing and recognizing expressions is the Face Action Coding System (FACS), which was developed by Friesen and Ekman (1978), and represents human facial muscle movements as a coding scheme. This method became very popular for emotion recognition systems, as shown in the review work of Cowie et al. (2001). One of the problems with methods that use the FACS is the amount of time necessary for extracting and recognizing the expressions, as debated by Sariyanidi, Gunes, and Cavallaro (2015), and the difficulty of representing spontaneous expressions, as discussed by Zeng, Pantic, Roisman, and Huang (2009).

To simulate a developmental-like emotional perception mechanism, we focus on the dimensional model-based representation. We train our CCCNN to learn strong and reliable emotion expression representations in different modalities and then replace the fully connected hidden and softmax layers of our network with a layer that implements Self-Organizing Maps (SOMs), introduced by Kohonen (1990). The SOMs are neural models where the neurons are trained in an unsupervised way to create a topological grid that represents the input stimulus. In a SOM, each neuron is trained to be a prototype of the input stimulus, meaning that after training, each neuron will have a strong emotional representation and neurons that are neighbors are related to similar expressions.

In our architecture, we usually implement a SOM with 40 neurons in each dimension. Empirically this was shown to be enough to represent up to 11 emotions for the Visual stream and up to 7 emotions using cross-modal representation. Figure 5 illustrates the updated version of our model.

3.1 Perception representation

After training, a SOM will create a grid of neurons, each one with the same dimensionality as the input stimulus. Analyzing a SOM is not an easy task, as stated by Vesanto (1999). The neurons of a SOM organize a projection of a high-dimensional data space into a set of neurons spread in a grid. That means that the knowledge of a SOM is represented by its topology. One way to interpret the neurons in a SOM is to use the U-Matrix, described by Ultsch (2003). The U-Matrix creates a visual representation of the distances between the neurons. Basically, you calculate the distance between adjacent neurons. The U-Matrix gives us a very important representation of the structural behavior of the SOM, in which we can identify different clusters of neurons. The U-Matrix of a SOM is defined as

$$\mathrm{U-Matrix} = \sum M_{m=1} \sum_{i=1}^{k} d(w - w_m) \qquad (7)$$

where M indexes the neighbor neurons, and w is the set of weights of each neuron. The distance calculation is given by $d(x, y)$, and is usually the Euclidean distance.

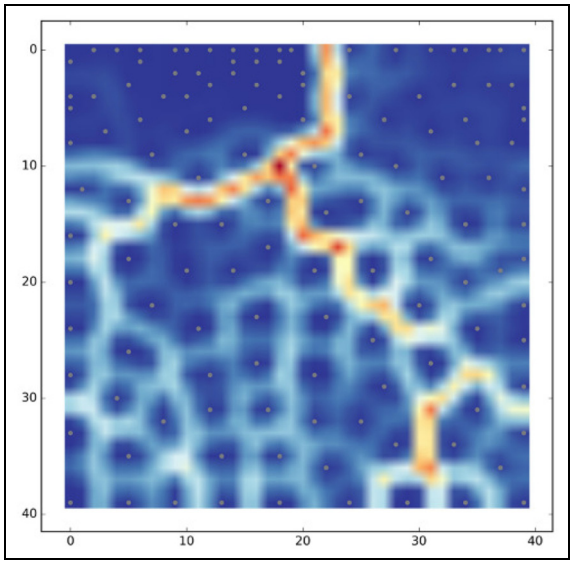

Figure 6. U-Matrix of a SOM with 40 neurons in each dimension and trained with happy, sad and neutral expressions. It is possible to see the neurons, represented by dots, in different regions, which represent the distances among the neurons.

After training, our SOM has neurons that represent emotion expressions, and we can visualize them by calculating the U-Matrix. Our SOM is trained completely unsupervised, which means that we do not identify the expressions we are showing to the network with any class and the U-Matrix shows the distribution of the neurons, or emotion expressions, over a grid. We use this grid to identify regions of neurons that have similar representation, and find certain patterns of the neuron distribution. Figure 6 illustrates an example of a U-Matrix calculated from a SOM with 40 neurons in each dimension and trained with three different expressions: happy, sad and neutral. It is possible to see the neurons, marked as the dots, and different regions based on the distances between the neurons.

The neurons that are strongly related to a presented input, will activate most: for instance, a certain neuron that activates for a happy expression will have a lower activation when a sad expression is presented. This way, by visualizing several activation maps, we can have an emotion representation that is very close to a dimensional perception, but learned in an unsupervised way. Figure 7 illustrates different activation maps. It is possible to see that the activation pattern changes when different happy, angry or neutral expressions are presented to the network.

The visualization of the knowledge learned by the SOM is not easy, similar to the human perception of emotions, as discussed by Hamlin (2013). She mentions that emotion expressions are learned by humans in a continuous process of perceiving new expressions and adapting them to previous knowledge. This process happens through the childhood by assimilating similar

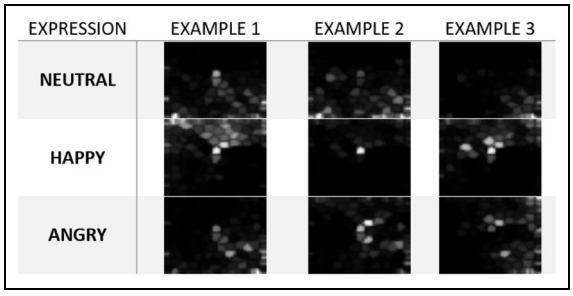

Figure 7. Examples of activation maps when three different expressions for each class are presented. It is possible to see that each class has an activation pattern different from the other classes.

emotions with known concepts, such as happiness, pain or solitude. That means that each person has their own emotion-perception mechanism, based on different features and different perceived emotions. Our model is inspired in this process, and first we use very strong feature representation, learned by the CCNN, to describe expressions. Later on, the learned expressions are clustered in similar concepts by our SOM, simulating the learning of new expressions. Our SOM also represents a unique perception representation, which could be related to a person's own perception.

3.2 Expression categorization

Using the regions of the U-Matrix we can create a categorical view of the network's representation. This helps us to use our model in emotion-recognition tasks. The advantage of using our model is that we can create different categorical models without re-training the network. If we want to analyze simple separations as positive and negative emotions, we can easily identify which regions of the network fire for these categories. If we want to increase the number of categories, we just have to increase the number of clusters. So, instead of finding regions that fire only for negative or positive, we can find regions that fire for happy, sad, surprised and disgusted.

To find these clusters, we use the U-Matrix to create a topological representation of the neurons and the K-means algorithm by MacQueen (1967) to cluster them. The K-means algorithm partitions a set of observations in N clusters, based on the distance from each observation to each other. The goal of the K-means is to minimize the within-cluster sum of squares, which could be defined as

$$K = \arg\min \sum_{i=1}^{k} \sum_{xES_i} ||(c - \mu_i)|| \tag{8}$$

where μ is the mean of each observation.

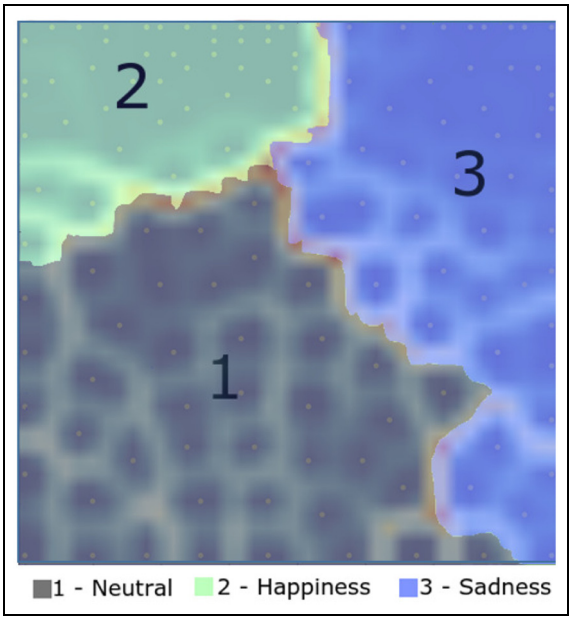

Figure 8. K-Means algorithm applied to the SOM illustrated at Figure 6. We cluster the neurons in three expressions: happy, sad and neutral. We use the K-means clusters to classify expressions.

The limitation of our model is directly related to the SOM architecture limitation: we have to define the number of neurons before training them, which restricts the number of expressions that can be categorized. However, with an arbitrary number of neurons, we can create different categories of expressions without retraining the network. The number of neurons depends directly on the amount and variety of expressions that were used to train the network, and the number of emotions to be categorized. Each prototype neuron learns an approximated representation of a set of expressions, so if the expressions used to train the SOM are not enough and the variety is too large, the network will need more neurons to represent emotional concepts.

Using the expression categorization, we can use our network to recognize different emotion categories. If, at first, we want to recognize only positive and negative emotions, we just have to define two clusters. Then, if we need to identify between a happy and an excited expression, we can apply the K-means algorithm only on the region of the network that has a bigger probability to activate for these concepts. In the same way, if we want to identify different kinds of happy expressions, we can create clusters only on this specific region. Figure 8 illustrates the application of K-means to the network illustrated in Figure 6. In this example, the network is clustered for three classes: happy, sad and neutral.

4 Experimental methodology

To evaluate our model, we perform three sets of experiments. In the first set we evaluate some aspects of the architecture: the impact of the input length and the use of the inhibitory fields. The second set of experiments evaluates the capability of the CCCNN to learn specific and crossmodal representations, and use them to classify emotion expressions. In the last set of experiments we evaluate our emotion-learning architecture, with the use of the SOM.

For all experiments, 30 experimental routines were performed and the mean of the accuracy was collected for each expression individually, which helps us to understand our model better.

4.1 Datasets

For our experiments we use four corpora. The first one is the FABO database, presented by Gunes and Piccardi (2006). This corpus is composed of recordings of the upper torso of different subjects while performing emotion expressions. This corpus contains a total of 11 expressions performed by 23 subjects of different nationalities. Each expression is performed in a spontaneous way, where no indication was given of how the subject must perform the expression. A total of 281 videos were recorded, each one having 2 to 4 of the following expressions: anger, anxiety, boredom, disgust, fear, happiness, surprise, puzzlement, sadness and uncertainty. Each expression starts with a neutral phase, and continues until the apex phase, where the expression is in its peak. We use the neutral phase for each expression to create a 12th neutral class in our experiments. Figure 9 illustrates images present in a sequence of an angry expression in the FABO corpus.

The second corpus is the SAVEE database, created by Haq and Jackson (2010). This corpus contains speech recordings from four male native English speakers. Each speaker reads sentences which are clustered into seven different classes: anger, disgust, fear, happiness, neutral, sadness and surprise. Each speaker recorded 120 utterances, with 30 neutral and 15 for each of the other emotions. All the texts are extracted from the TIMIT dataset and are phonetically balanced. Each recording contains the audio and face of the participant, with facial markers. The markers are present to be used for systems that need them, and unfortunately we cannot remove them from the image. Figure 10 illustrates faces of a subject while performing an angry expression in the SAVEE corpus.

The third corpus is the EmotiW database, published by Dhall, Goecke, Lucey, and Gedeon (2012). This corpus contains video clips extracted from random movies and separated in seven classes: anger, disgust, fear, happiness, neutral, sadness and surprise. A total of 1000 videos with different lengths are available, separated into training and validation sets. The test set is available, but without any labels, and includes 700 extra videos. Therefore, we only evaluate our model on the validation set. This challenge is recognized as one of the

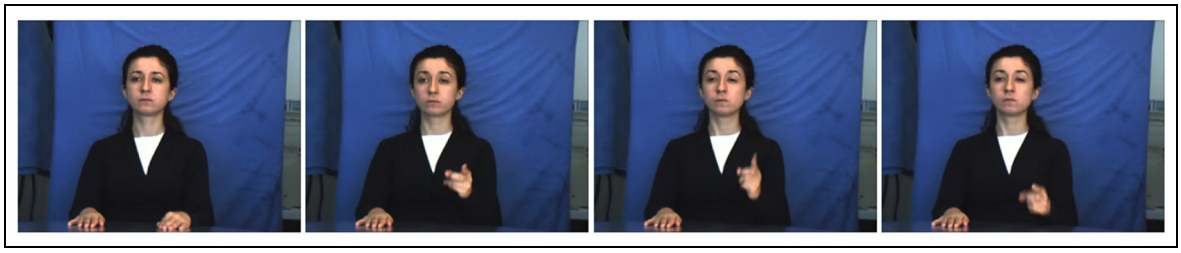

Figure 9. Examples of images with an a angry expression in the FABO corpus.

Figure 10. Faces with an angry expression in the SAVEE corpus.

Figure 11. Example of an angry sequence in the EmotiW corpus.

most difficult tasks for emotion recognition, because the movie scenes contain very cluttered environments, occluded faces, speech, music, sound effects, more than one speaker and even animals. Figure 11 illustrates some frames of an angry expression in the EmotiW corpus.

One extra corpus is used to train the Music stream of the auditory information. The GTZAN corpus by Tzanetakis and Cook (2002) is not directly related to emotion recognition, but to music genre classification. The task of music genre classification is similar to music emotion classification by Kim, Valitutti, and Calvo (2010), because the idea is to cluster audio segments that are closely related based on auditory features. Music genres can also be used for emotion classification, since, for example, blues and soul music are more related to sadness or lonely feelings, and pop music more to happiness, see Kim et al. (2010). This database contains 1000 songs, each one 30 s with a sampling rate of 22050 Hz at 16 bit, divided into 10 musical genres: blues, classical, country, disco, hip hop, jazz, metal, pop, reggae and rock.

All the experiments were performed by using four-fold cross validation, except for the EmotiW. This corpus has a pre-defined separation for the testing and validation sets, and we do not use the test set in most of the experiments due to the lack of labels.

4.2 Architecture experiments

According to Ekman (2007), an expression occurs between 300 ms and 2 s. To evaluate an optimal approximation of sequence length, we evaluated our Face channel with four different input lengths: 40 ms, 300 ms, 600 ms and 1 s. For this experiment we use the FABO corpus, and as the sequences in this corpus were recorded with a frame rate of 30 f/s, that means that we evaluate the use of 1, 9, 18 and 36 frames as input. We also evaluated the input length for the Movement channel. First, we evaluate the use of 2 frames to compose a movement representation, then 5 frames, 10 frames and lastly 15 frames. This leads to feeding the network with 15, 6, 3 and 2 movement images, respectively.

We then evaluate the use of the inhibitory fields on the Visual stream, by applying them in different layers. We show how the inhibitory fields affect each representation of each layer and why we only use them on our Face channel.

For the auditory representation, we follow indications in the work of Abdel-Hamid et al. (2014) for the Speech channel and George and Shamir (2015) for the Music channel. Separating the same 1 s of representation and using the window and sliding values indicated in this work produced the best results, so we kept them. Also, the use of inhibitory fields on the auditory channel did not produce any improvement in the results, causing exactly the opposite: an overfitting of the filters made the network lose the focus completely during training.

4.3 FABO Experiments

Using the FABO corpus we evaluate the Visual stream of the network. In this set of experiments, we evaluate the use of the Face and Movement channels individually, and then both of them at the same time.

With this experiments we show in detail the impact that each modality has in different expressions. As the FABO corpus deals with secondary expressions, it is possible to see how our visual representation behaves for very different expressions, such as happiness and sadness, or very similar ones, such as boredom and puzzlement.

4.4 SAVEE Experiments

The SAVEE corpus was used to evaluate the Auditory stream of the network. As this corpus also has visual information, with the recording of the faces of the subjects, we also evaluate the Face channel and the crossmodal representation obtained with the use of the auditory channels and the Face channel.

The audio data of the SAVEE corpus contains only speech and no music. This way, we evaluate the use of only the Speech channel but also the use of the Music channel, pre-trained with the GTZAN corpus, integrated with the Speech channel. This way, we can show how the model behaves with specific training, but also, how the integrated model, which has speech- and music-specific representation, behaves when only one type of input is present.

4.5 EmotiW Experiments

The EmotiW corpus contains the most complex emotion expressions in our experiments. The video clips contain different subjects (sometimes at the same time), music, speech, different illumination in the same scene and various face positions. This makes the emotion classification in this dataset very difficult.

We evaluate the use of our channels trained with this dataset; first each channel individually and then the integration of visual-only streams and auditory-only streams. Finally, we evaluate the audio–visual representation. Each of these experiments is performed with two different training strategies: one with, and one without the pre-training of the filters. We use the FABO corpus to pre-train the filters of the Visual stream and the SAVEE and GTZAN corpus to pre-train the Auditory stream.

The auditory information of this corpus is not separated, and thus we feed our Auditory stream with the same input data. The idea here is that the Auditory stream will represent mostly the speech information in one channel, and the music/background sound in the other.

All of the results are compared and we show for the six basic emotions, plus a neutral category, how each of the modalities behaves, and the advantage of using the pre-training strategy.

4.6 Expression learning experiments

For the expression learning experiments, we use the trained filters of the CCCNN to extract high-level expression representations and train a SOM with them. After training, the SOM is used in classification tasks, by using K-means to cluster the neurons in a number of specified classes. We compare the use of the SOM with the CCCNN performance for classifying crossmodal data with the EmotiW corpus. These experiments show the capability of the SOM to generalize expressions.

We also measure the capability of the SOM to learn new expressions. For that, we train a SOM with a limited set of expressions, with sad and happy emotions being performed. Then, we systematically present new expressions to the SOM, such as anger, disgust and surprise, and we show the mean of the activation maps for each expression. This way we show the capability of the SOM to learn different expressions. For these experiments we use the FABO corpus, because it contains a controllable environment, which is not present on the EmotiW dataset.

In the last round of experiments, we show the use of the SOM for analyzing the behavior of expressions. We perform experiments with the SAVEE corpus only, which contains data from four different subjects. We train one SOM for each subject and compare the differences of the expressions based on the clusters of each SOM.

5 Results

5.1 Architecture experiments

The results obtained when training the Face channel with different sequence lengths showed that the use of nine frames produced the best results, as exhibited in

Table 1. Reported accuracy, as percentages, for different lengths of the input sequence, in frames, for the Face channel, and in movement representations for the Movement channel trained with the FABO corpus.

Face channel				
Sequence length	1	9	18	36
Accuracy(%)	64.8	*80.6*	73.6	49.4
Movement channel				
Motion representations	15	6	3	2
Accuracy(%)	48.3	67.9	*74.8*	66.2

Table 2. Reported accuracy, as percentages, for the use of inhibitory neurons in different layers of the Face and Movement channels trained with the FABO corpus.

Face channel					
Layers	None	L1	L2	All	
Accuracy(%)	80.6	59.9	*87.3*	64.4	
Movement channel					
Layers	None	L1	L2	L3	All
Accuracy(%)	*74.8*	41.3	47.8	48.8	45.8

Note: The italic text indicate the highest values.

Table 1. As the FABO corpus was recorded with 30 f/s, the use of 9 frames means that the sequence has an approximate length of 300 ms. A sequence with this length is congruent with the description of face expressions. The use of longer expressions, with 1 s, produced the weakest results.

The Movement channel receives as input a sequence of motion representations of 1 s of the expression. This means that each representation of this sequence is composed of several frames. The results, exhibited in Table 1, show that the use of 3 movement representations obtained the best performance, meaning that each movement representation is composed of 10 frames and each motion representation captures 300 ms. Our results show that using a minimum number of frames to capture the movement, 2 frames per motion representation and 15 frames as the channel's input, produces the worst result.

In the second round of experiments, we evaluate the use of inhibitory neurons in our visual channels. We evaluate the use of the inhibitory fields on each of the layers, and in combination on all layers of each channel. Table 2 exhibits the results. The application of the inhibitory fields on the Movement channel did not produce better results, due to the fact that the movement representation is already a specified stimulus, and the filters alone were capable of coping with the complexity of the representation. It is possible to see that when applied to the last layer of the Face channel, the inhibitory fields produced better results, confirming that the strong extra-specification on the last layer is beneficial for face expression recognition.

Table 3. Reported accuracy, as percentages, for the Visual stream channels trained with the FABO corpus. The results are for the Face channel (F), Movement channel (M) and the integrated Face and Movement channel (FM), representing the Visual stream (V).

Class	F	M	FM
Anger	74.5	66.3	95.9
Anxiety	78.6	80.5	91.2
Uncertainty	82.3	75.8	86.4
Boredom	93.4	76.3	92.3
Disgust	78.3	65.9	93.2
Fear	96.3	80.0	94.7
Happiness	93.7	60.3	98.8
Negative surprise	67.2	32.4	99.6
Positive surprise	85.7	65.7	89.6
Puzzlement	85.4	84.8	88.7
Sadness	89.6	80.1	99.8
Mean	87.3	74.8	93.65

Table 4. Comparison of the accuracy, as a percentage, of our model with state-of-the-art approaches reported with the FABO corpus for representations of face, movement, and both integrated.

Approach	Face	Movement	Both
Barros et al. (2015a)	72.7	57.8	91.3
Chen et al. (2013)	66.5	66.7	75.0
Gunes & Picccardi (2009)	32.49	76.0	82.5
CCCNN	87.3	74.8	93.65

5.2 FABO Experiments

The combined results of the FABO experiments are exhibited in Table 3. It is possible to see that overall, the mean accuracy of the integrated representation is the highest. Also, it is possible to see how some expressions behave with different modalities. For example, anxiety and puzzlement expressions had a performance similar to the Face and Movement channels alone, but increased when the integrated representation was used. Also, there was a great increase in the performance for disgust and negative surprise, showing that for these expressions the integrated representation provided more information than each modality individually.

Comparing our model with state-of-the-art approaches using the FABO corpus shows that our network performed similar, and in the Face representation better. Table 4 exhibits this comparison. The works of Chen et al. (2013) and Gunes and Piccardi (2009)

extract several landmark features from the face, and diverse movement descriptors for the body movement. They create a huge feature descriptor for each modality, and use techniques such as SVM and Random Forest, respectively, for classification. It is possible to see that the fusion of both modalities improved their results, but the performance is still lower than ours. In previous work, we used a Multichannel Convolution Neural Network (MCCNN), published as Barros et al. (2015a), to extract facial and movement features. This network produces a joint representation, but our current CCCNN improved this representation with the use of separated channels per modality and the application of inhibitory fields. It is possible to see a substantial improvement on the movement representation, mostly because we use a different movement representation in the CCCNN.

5.3 SAVEE Experiments

The results with the SAVEE experiments are exhibited in Table 5. It is possible to see that the auditory information obtained the lowest accuracy, and among them the pre-trained representation was the one with the lowest general accuracy. This occurs because the data in the SAVEE corpus does not contain music, only speech, which reflects directly on the performance obtained by the network. Still, it is possible to see that the auditory channel composed of the Speech and Music does not substantially decrease the performance of the network, but makes it more robust to deal with speech and music data.

We also see that the face representation obtained a similar performance to the auditory one, but when combined, the performance tends to increase. This is due to the fact that when both, face and auditory information, are present, the network can distinguish better between the expressions. This is demonstrated by the performance of the model for anger, sadness and surprise, which have a similar performance in individual channels and a higher one in the integrated representation.

Our approach showed to be competitive when evaluated with the SAVEE corpus. When compared to state-of-the-art approaches, our representations showed a result comparable with the work of Banda and Robinson (2011). They use a decision-based fusion framework to infer emotion from audio–visual inputs. They process each modality differently, using linear binary patterns to represent the face expressions and a series of audio features to represent speech. After that, an in-pairs SVM strategy is used to train the representations. Our network has a similar performance for face representation, but a higher accuracy for audio. We improved by more than 10% the accuracy of the speech representation. For the multimodal integration, our network has been shown to be competitive, and performed similarly, but with a much less costly feature representation process. The authors of the SAVEE dataset, Haq, Jackson, and Edge (2009), also did a study to examine the human performance for the same task. Using the same protocol, a 4-fold cross validation, they evaluated the performance of 10 subjects on the recognition of emotions on the audio and video data. The results showed that most approaches exceeded human performance on this dataset. This happens for most of the compared methods, and the probable cause is that the methods create a very specific representation of the expressions (only the six basic emotions plus neutral), while humans have a larger amount of learned representations, which can help when determining unknown and spontaneous expressions, but could hinder recognition in restricted scenarios. Table 6 exhibits the state-of-art results and human performance on the SAVEE dataset.

Table 5. Reported accuracy, as percentages, for the Auditory and Visual stream channels trained with the SAVEE corpus. The results are for the Face channel (F), Speech channel (S), Speech and pre-trained Music channel, representing the Auditory stream (A) and the integrated audio–visual streams, with the Face, Speech and Music channels (AV).

Class	F	S	A	AV
Anger	95.4	95.0	92.6	100
Disgust	95.6	100	88.0	100
Fear	89.7	88.0	85.5	100
Happiness	100	81.1	86.1	95.0
Neutral	100	100	91.3	100
Sadness	90.0	93.5	87.4	96.5
Surprise	86.7	86.5	80.5	96.7
Mean	93.9	92.0	87.3	98.3

5.4 EmotiW Experiments

The EmotiW corpus proved to be a very difficult challenge. Table 7 illustrates all the results on the corpus. It is possible to see that the visual representations, represented by the columns F, M and V, reached better results than the auditory representations, presented in columns S, Mu and A.

The visual representations presented a very interesting distribution of accuracies. It is possible to see that when the expressions were represented by the

Table 6. Performance of state-of-the-art approaches on the SAVEE dataset.

Methodology	Face	Audio	Both
Banda& Robinson (2011)	95.0	79.0	98.0
Haq et al. (2009)	95.4	56.3	97.5
CCCNN	93.9	92.0	98.31
Human performance	88.0	66.5	91.8

Table 7. Reported accuracy, as percentages, for the Auditory and Visual stream channels trained with the validation set of the EmotiW corpus. The results are for the Face channel (F), Movement channel (M), Face and Movement channel together, representing the Visual stream (V), Speech channel (S), Music channel (Mu), Speech and Music channel together, representing the Auditory stream (A) and visual–auditory integration (AV).

Class	F	M	V	S	Mu	A	AV
Anger	70.2	50.8	77.8	56.4	50.7	70.1	80.3
Disgust	18.2	9.4	18.7	12.4	2.6	15.2	23.4
Fear	21.4	16.8	20.2	7.8	6.5	7.2	30.8
Happiness	67.2	75.6	77.8	59.1	65.4	72.0	81.2
Neutral	67.2	57.7	70.9	10.8	15.6	25.4	68.7
Sadness	22.4	21.2	23.2	8.3	9.8	16.2	24.5
Surprise	5.4	10.0	12.1	0.0	2.1	4.1	14.0
Mean	38.8	34.5	42.9	22.1	21.8	30.0	46.1

movement, column M, happy and sad expressions performed better than the others, showing that for happy and sad expressions the movements were more reliable than the face expression itself. When integrated, the visual representation improved the performance of most expressions, in particular, surprised, angry and happy expressions, which indicates that these expressions are better recognized when movement and face expressions are taken in consideration.

The auditory representation indicates that most of the expressions are not well recognized with auditory information only, exceptions are angry and happy emotions. This can be related to the nature of the dataset, because usually in movies happy and angry are expressed with similar song tracks or intonations. The integrated representation for the Auditory stream performed better than the individual ones in all the expressions.

Finally, the multimodal representation was the one with the best performance. We see an improvement in sad and angry expressions, but also in fear and surprised ones. This is due to the fact that the combination of different soundtracks, facial expressions and movement for these expressions represents them better than a single modality. In general, it is possible to see that surprised, disgusted and sad expressions were the ones with the lowest performance in all modalities.

Table 8 exhibits the results on the EmotiW dataset. On this dataset, the performance of our model dropped, but as Table 8 shows, this is also a much harder task. Due to the variability of the data, neither of the modalities provides an overall high accuracy. Our model results are competitive with the state-of-the-art approaches, and performed better than the baseline values for the competition. The works of Liu et al. (2014) and Kahou et al. (2013) extract more than 100 auditory features each, and use several CNNs to extract facial features. They feed a vector composed of

Table 8. 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	Video	Audio	Both
Liu et al. (2014)	45.28	30.73	48.53
Kahou et al. (2013)	38.1	29.3	41.1
Dhall et al. (2014)	33.15	26.10	28.19
CCCNN	42.9	30.0	46.1

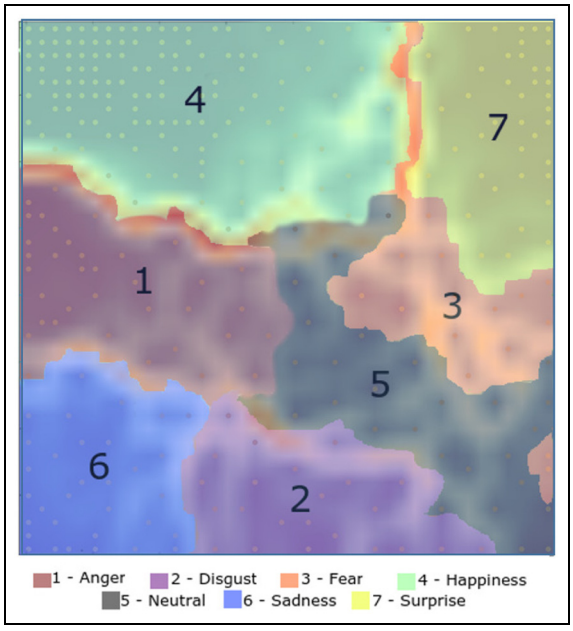

Figure 12. K-Means algorithm applied to the SOM trained with the EmotiW multimodal representation. Six emotions were clustered: surprise, sadness, anger, happiness, fear, neutral and disgust.

the output of the CNNs and the auditory features into several classifiers such as SVM or multilayer perceptrons to classify them. Our model results show that we can actually obtain similar generalization capability using a simple and direct pre-training strategy without the necessity of relying on several different feature representations.

5.5 Emotion categorization

For these experiments, we trained our SOM with the emotion representation obtained by the CCCNN of the previous experiment. We then cluster the neurons of the SOM in 7 regions with a K-means algorithm, so each region represents one class of the EmotiW corpus. Figure 12 illustrates the clustered regions from 0 to 6, respectively: anger, disgust, fear, happiness, neutral, sadness and surprise. It is possible to see that the neutral expressions, represented by class number 5, have as neighbor almost all the other expressions. Also, angry

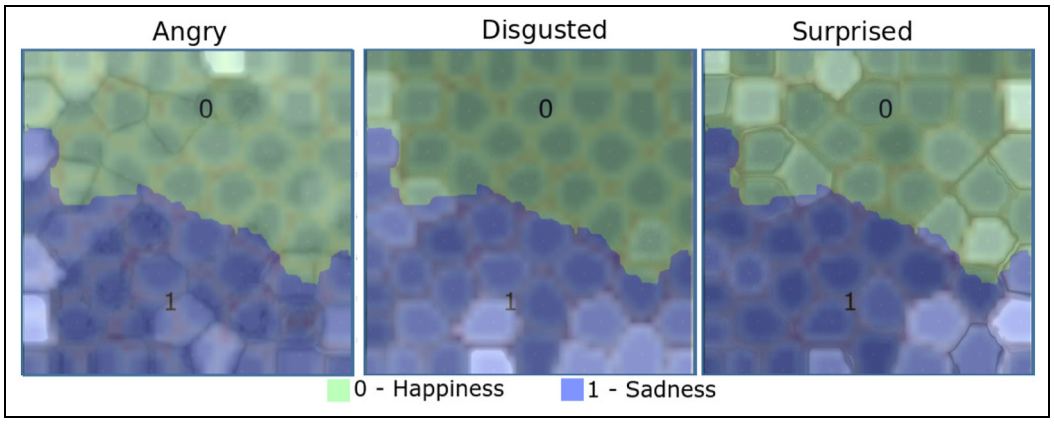

Figure 13. Activations plotted on top of a clustered SOM. The SOM was trained with sad and angry expressions and each activation shows the mean activation map when feeding the network with angry, disgusted and surprised expressions.

Table 9. Reported accuracy, as percentages, for the multimodal representation in the validation set of the EmotiW corpus. The results are for the CCCNN and the SOM.

Class	CCCNN	SOM
Anger	80.3	85.3
Disgust	23.4	30.3
Fear	30.8	32.1
Happiness	81.2	82.3
Neutral	68.7	67.3
Sadness	24.5	31.7
Surprise	14.0	17.6
Mean	46.1	49.5

expressions, class number 1, are between happy, class number 4, and sad expressions, class number 6. And lastly, it is possible to see that fear expressions, class number 3, are closely related to surprise expressions, class number 7. In this case, some of the fear expressions are between happy and surprise.

Using the clusters we calculated the accuracy of the SOM in the validation set of the EmotiW corpus. Table 9 exhibits the results. It is possible to see that with the SOM clustering, expressions such as disgust and sadness show an increase of almost 7% in performance. As we see in the cluster, sad and disgusted expressions are neighboring regions, and the application of the SOM created a better separation border, which would explain the performance increase. In general we have an improvement of more than 3% in the accuracy when using the SOM.

5.6 Learning new emotions

In our next experiment, we trained the SOM with happy and sad expressions from the FABO corpus. We then proceed by feeding to the network angry, disgusted and surprised expressions, and generate the mean of the activation maps for each set of expressions. Figure 13 illustrates the activations for each new set of expressions plotted on top of the clustered SOM. In this experiment, the network never saw angry, disgusted or surprised expressions and we can see how the neurons activate when these expressions are presented.

Angry expressions activated a mixed region of neurons, between the sad and happy regions. Two neurons had a higher activation, in both regions. This is congruent with the regions found when analyzing the EmotiW SOM, where angry expressions were represented between happy and sad. Disgusted expressions were mostly activated by neurons on the sad region, which is also congruent with the cluster of the EmotiW SOM. And finally, the surprised expressions were mostly activated in the happy regions, with some activation in the angry region.

We then proceeded to re-train the network with the new expression. We used the network trained with sad and happy expressions, and created four new networks, three of them trained with the addition of one new expression, and the fourth one with all five expressions. Figure 14 illustrates the clusters of each network. We can see that the disposition of the new clusters is similar to the activation maps of the network trained with only two expressions. This demonstrates how each emotion expression can be related to others, and our network is able to use this relation to learn new expressions.

5.7 Expression behavior

In the final experiments with the SOM, we trained one SOM with expressions, represented by Face and Speech channels, from each one of the four subjects on the SAVEE corpus, which are identified as DC, JE, JK and Kl. We trained each SOM using a four-fold cross validation strategy, only with the data of each individual subject. We then calculated the accuracy for each subject, which is exhibited in Table 10.

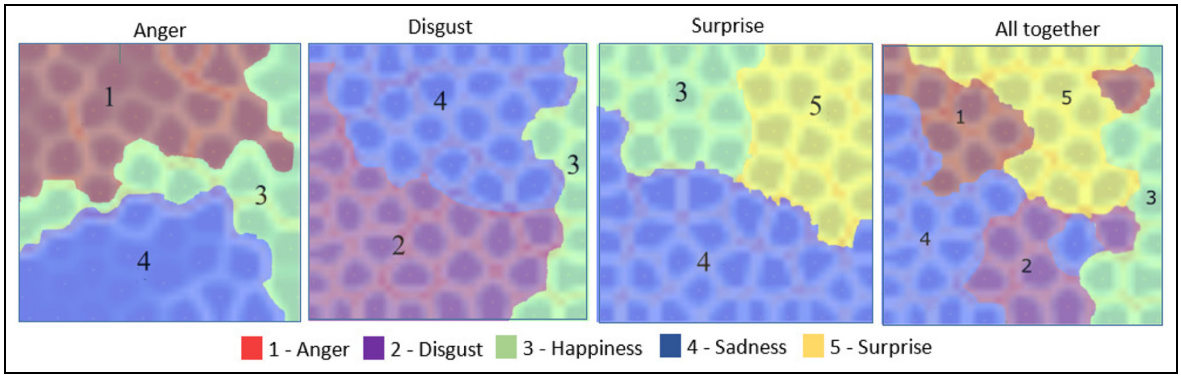

Figure 14. We train a network with two kinds of expressions: happy and sad. Systematically we add one different expression and re-train the network. At the end, we train a network with the five expressions together.

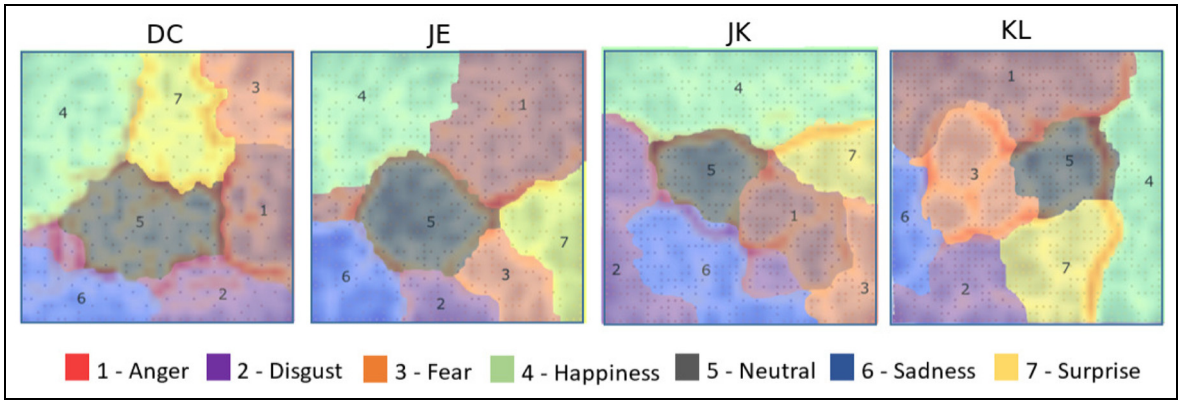

Figure 15. Trained networks with expressions of each subject of the SAVEE corpus. It is possible to visualize how differently each subject expresses by analyzing the network clusters.

Table 10. Reported accuracy, as percentages, for the Auditory and Visual stream channels trained with a SOM and the SAVEE corpus per subject.

Class	DC	JE	JK	KL
Anger	100.0	94.3	100.0	92.0
Disgust	100.0	100.0	100.0	90.9
Fear	100.0	100.0	96.7	100.
Happiness	99.4	99.1	100.	97.7
Neutral	98.3	100.0	100.0	96.7
Sadness	96.7	97.8	100.0	97.8
Surprise	100.0	100.0	97.9	98.2
Mean	99.1	98.7	99.2	98.3

We separated the regions of each SOM into seven classes, and produced cluster images for each subject, which are illustrated in Figure 15. Analyzing each cluster, we can see that the same expressions have different regions for each subject. Analyzing these images, it is possible to obtain some information about how each subject expresses itself. For each subject, the same number of samples is recorded for each emotion category, so there is no bias to one expression in each subject.

Except for the network of subject JE, all others clustered surprised expressions in a neighbor region to happy expressions. On other hand, all of them clustered surprise in a neighbor region to fear expressions. That indicates that JE surprised expressions are less happy than the others. Also, the disgust expression is different for each subject. Although all of them have disgusted expressions as a neighbor of sad expressions, the other neighbors change. It is possible to see that for DC, disgusted expressions are closely related to angry ones, for JE with fear, JK with happy and KL with surprised expressions. Looking for the region that each expression takes part in, it is possible to see that JK's network clustered happy expressions with a larger region than the others, which could be an indication that the happy expressions in JK are more different within each other than the others. The same happens with JK's disgusted expressions. On the other hand, his neutral expressions have a smaller region than the others, indicating that

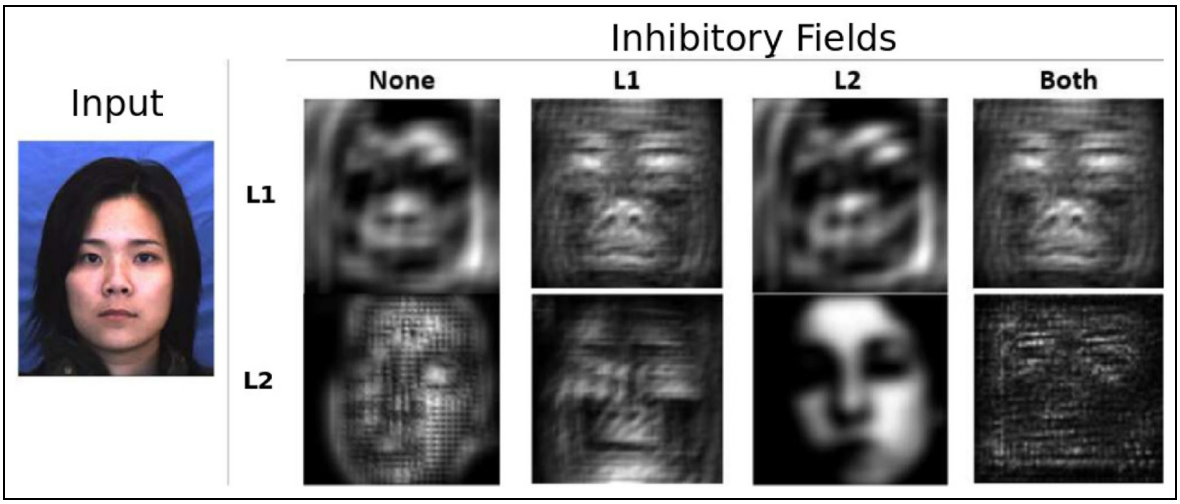

Figure 16. Implementing inhibitory fields in different layers of the network produces different features. Each visualization corresponds to one filter on a determined layer. It is possible to see how the inhibitory fields affect the feature extraction capabilities of each layer.

most of his neutral expressions are very similar to one another.

6 Discussion

In this section we discuss three concepts. First we analyze the CCCNN architecture, and how the introduction of inhibitory fields and the crosschannel contribute to the expression representation. Second, we discuss how the model represents multimodal stimuli, how the expression is decomposed inside the model, and what each layer represents. Lastly, we discuss the role of the SOM in learning similar expressions, and we associate this mechanism to a concept of emotional neurons.

6.1 Inhibitory fields and crosschannels

The application of inhibitory fields has been shown to increase the performance of the network only when they were implemented in the last layer of the face channel. That was caused by the overfitting that the inhibitory fields produced in the layer's filters. When the inhibitory fields were applied to the first layer, the filters learned more complex patterns, which did not help in the feature generalization. That phenomenon is easily visible when we visualize the features that the network learned using the deconvolution process illustrated in Figure 16, which shows the visualizations of the internal knowledge of one filter in different layers of the network.

When no inhibitory filter was implemented, it is possible to see that in the first layer the network learned some edge detectors, which could filter mostly the background and hair of the person. In the second layer, the network constructed a higher level of abstraction, mostly the shape of the face, and some regions such as eyes, mouth and nose are roughly highlighted. When we implemented the inhibitory fields in the first layer only, we found that more information was filtered. The filters detected more precise regions, filtering much more information that is relevant to represent the facial expression. This caused a problem in the second layer, which then tried to learn very specified concepts, and constructed a very limited representation. When the inhibitory fields were applied in the last layer, we found a very clear distinction in the representation. The shape of the face is very clear, but regions such as eyes, nose and mouth are better represented when no inhibitory fields are applied. Finally, when we applied the inhibitory fields in both layers, the final representation does not contain any reliable information with some very rough representation of the eyes and nose.

The crosschannels also have an impact on the quality of the extracted filters. Our crosschannels integrate two channels into one representation, which was shown to be more efficient and robust, but also reduced the dimensionality of the data. The application of the crosschannels created a new representation of the input stimulus, which is different from the individual representation. Figure 17 illustrates the visualizations of the last layer of the individual channels and the crosschannel. We can see that the crosschannel features are different from the individual representation, and they changed to capture an important feature: hands over the face. Furthermore, we see that the facial features changed drastically to incorporate the movement of the hands, which are now also highlighted in the movement channel.

6.2 Expression representation

The application of the visualizations also helps us to understand how the network represents an expression.

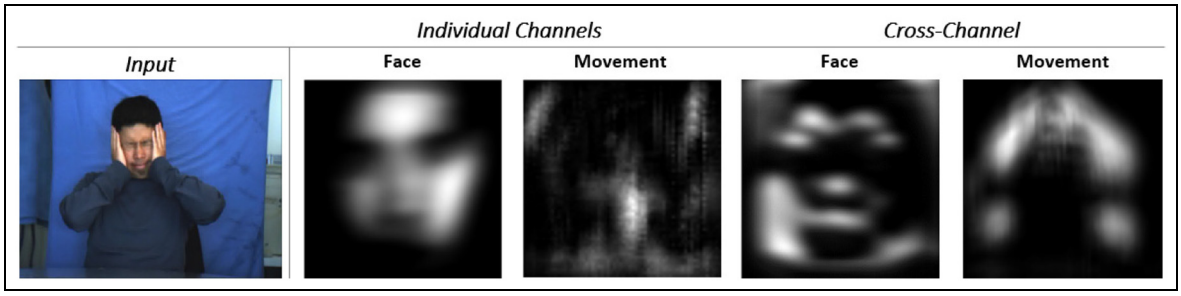

Figure 17. Applying the crosschannel on the individual representations brings results on different features. Note that the face representation after the application of the crosschannel changed to include the hand movement.

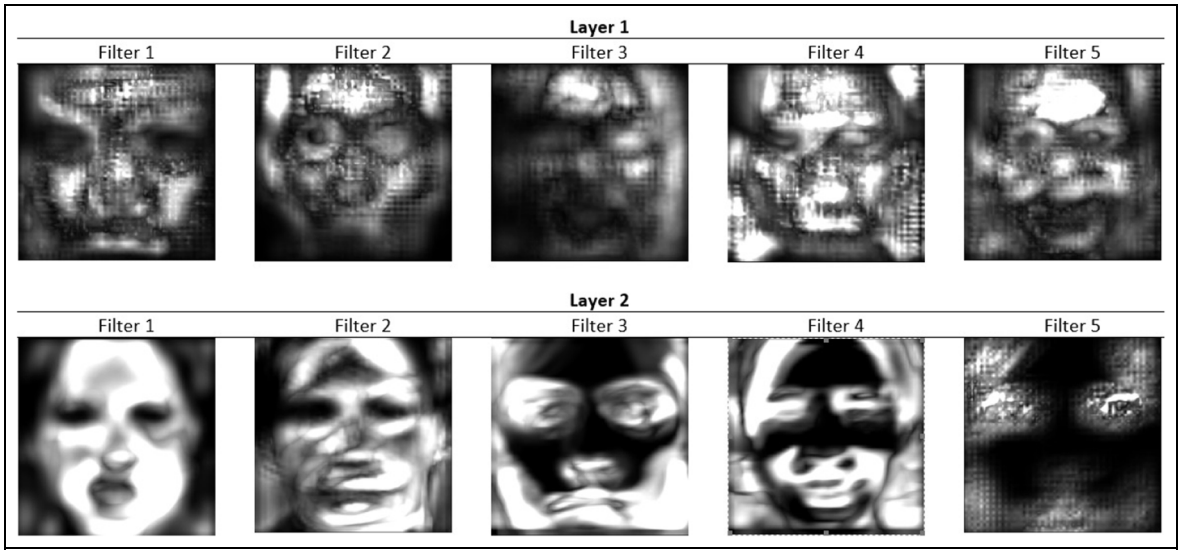

Figure 18. Mean visualization from all images in the FABO corpus per filter in all the layers of the Face channel. It is possible to use specialized filters, which help us to understand how the expression representation is created.

It is possible to see how the expressions are decomposed inside the network, and gain an insight into the role of each layer of the network in building the expression representation. By visualizing the same region of neurons for several images, it is possible to identify for which regions those neurons activate most. This way, we can analyze which parts of the input stimulus activate each filter of the network. Figure 18 illustrates this concept, where it is possible to see what each filter codes for in each layer. To generate these visualizations, we created an average per filter in the Face channel for all the images in the FABO corpus.

The filters learn to represent different things, which are complementary for the emotion expression. In the first layer, mostly background and hair information are filtered. Filter 5 highlights the region of the mouth out of the image, while filter 2 keeps the eye information. The most interesting representations occur in the second layer, where filters 1 and 2 represent mostly the face shape and positions of eyes, nose and mouth. Filters 3 and 4 represent the eyes, nose and mouth shapes, where filter 3 activates mostly for the cheeks and closed mouths and filter 4 for opened mouths. Different from the others, filter 5 specialized mostly in eyebrows.

Our network filters react to very specific patterns on the input images, which are related to human facial expressions. We can see how these patterns are strong when we send to the network, images that resemble human expressions, illustrated in Figure 19. The network highlighted regions that were closely related to human features. In the image with the dog, the position of the eyes and mouth were detected, and in the Don Quixote painting, the shape of the face was highlighted. In all images, it is possible to see that the filters of the network highlighted regions of interest that have a similar contrast to some facial features, such as face, eyes, and mouth shapes. On the other hand, the network is strongly domain-restricted. It will always try to find human facial features in the images, even when they are not present. This can cause problems, especially in the EmotiW corpus, illustrated in the last column of Figure 19.

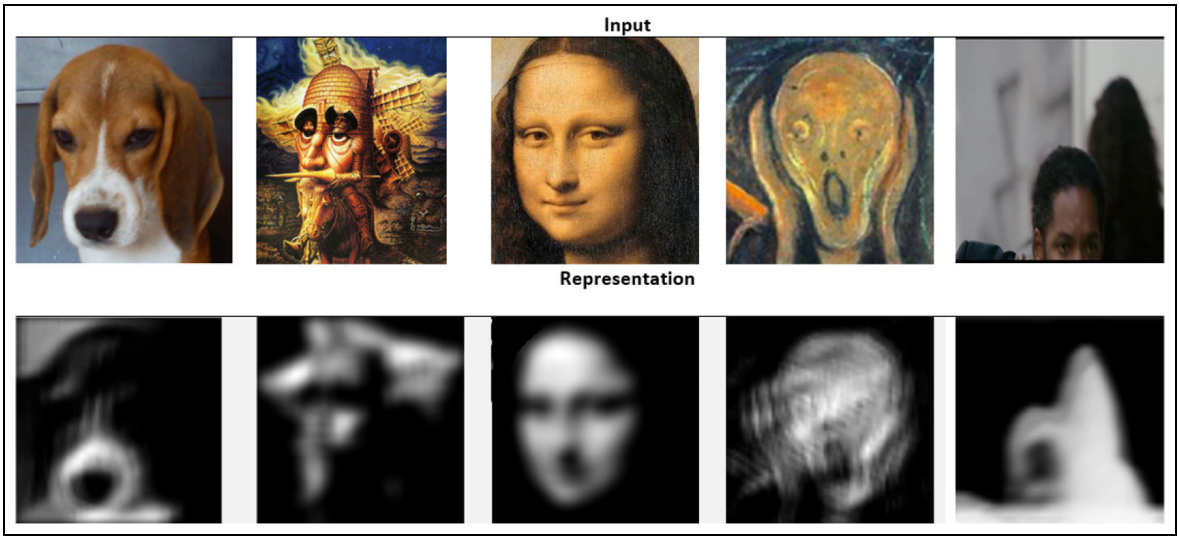

Figure 19. Visualization of the facial representation for different images. We see that the network tries to find human facial features, such as mouths, eyes and face shapes in the images.

6.3 Affective neural networks

Some regions of the CCCNN code for specific features, such as face shape, eyes, mouth among others. However, once these features are related to an emotion expression, it is the responsibility of the fully connected hidden, and softmax layers to classify these features into emotions. These layers do not store any information about the expressions themselves, only about the separation space. Replacing these neurons by a SOM gives the model a powerful tool to represent emotion expressions. Besides creating a more flexible separation region, the SOM allows the model itself to store information about the expressions.

Each neuron in the SOM represents a prototype of an expression, which is tuned to be similar to the data used to train the model. This means that each neuron alone codes for an expression, and neighbor neurons code similar expressions. In this way, we can simulate the spatial separation that the hidden and the softmax layers produce by clustering the neurons in different regions, giving the SOM the capability to classify expressions. This means that a real expression has to be represented by one prototype expression in order to be classified, which improved the performance of classification tasks.

The prototype expressions also help our model to code the concept of the expression itself. While the filters on the CCCNN code for specific features from the input stimulus, each group of neurons in the SOM code for similar expressions, giving our model a complete representation of the emotional expression, from the input stimulus to the expression representation itself. This idea differs from most of the work in the area, which learns how to represent features or how to create a separation space to classify these features into known expressions.

We can actually use the visualizations to gain an insight into what expressions the model learns. When visualizing an input, we backpropagate the responses that the input produced in our filters, however, by using the prototype neuron representation instead of the image representation, we can visualize which expression this neuron learned. By doing that for several images and several neurons, we can actually identify how these expressions change through the network, which helps us to understand the clusters of the SOM and the network representation itself.

Taking as an example the network trained for each subject of the SAVEE corpus, we can visualize the expressions learned by each neuron. Figure 20 illustrates some neurons of two subjects that are in the same region and correspond to angry expressions. It is possible to see that both networks have different representations for angry expressions, depending where the neurons are. In DC, it is possible to see that an expression closer to the fear region, produces a different mouth shape to the one closer to the surprise region. And for JE it is possible to see that all three representations have different eye and mouth shapes.

7 Conclusion and future work

We propose a novel architecture for emotion expression representation and learning. Our model implements CCCNNs to learn specific features of audio–visual stimuli. The network implements several channels, each one learns different features from each modality and applies a crossconvolution learning scheme to generate auditory and visual representations of emotion expressions.

On top of the CCCNN filters, we implement a SOM layer, which is responsible for learning how to separate

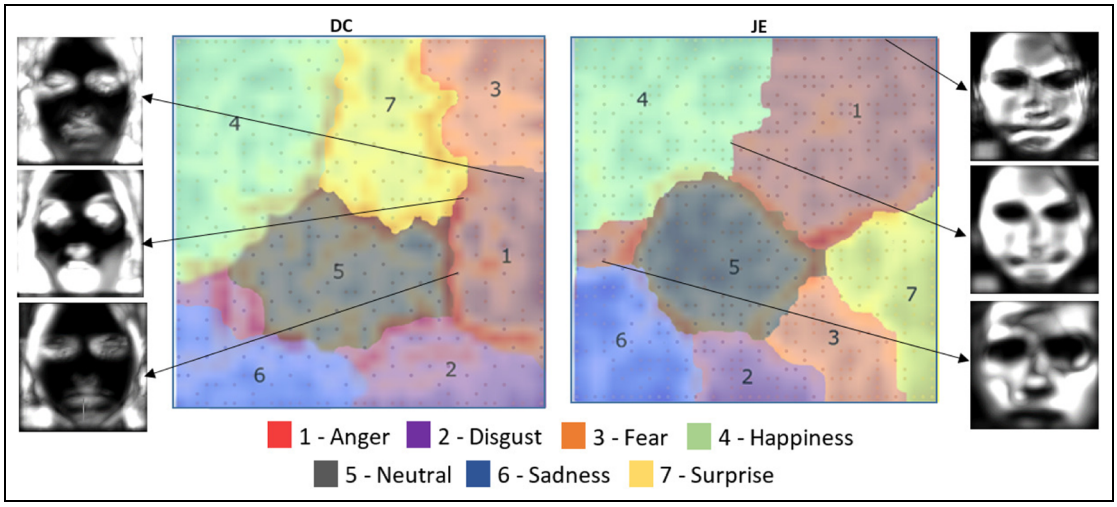

Figure 20. Visualization of the neural emotional representation for two subjects of the SAVEE corpus. It is possible to see how neurons that are closer to different regions of the network, store different expressions.

the representations into expressions. The SOM creates a series of prototype neurons, each one of them coding an expression representation. This means that neighboring neurons have a similar expression representation, and when we cluster these neurons in regions, we can obtain emotion expression categories.

To evaluate our model, we use three different corpora: the FABO database with visual expressions, the SAVEE database with audio–visual expressions, and the EmotiW database, which contains audio–visual clips extracted from different movies. Each corpus contains different expression information, and we use them to fine-tune the training of our CCCNN and to evaluate each individual modality. Our network was shown to be competitive, and in the case of the FABO corpus, better when compared to state-of-the-art approaches.

We also introduce mechanisms that allow us to understand and identify the knowledge of the network. By using the deconvolution process to visualize the internal representation of the CCCNN filters and the K-mean cluster algorithm to identify regions in the SOM, we showed that our model has a very wide emotion expression representation. We can use our model to classify emotions in categories, or in a dimensional space. Also, our model is suited for learning new expressions and we demonstrate its capability to help to understand emotion behaviors.

One of the limitations of our model is the SOM itself, which is limited by the number of neurons in its grid, which means that at some point the number of neurons will not be enough to represent new expressions and some of the old expressions will be forgotten. To overcome this limitation we will introduce the use of Growing-When-Required networks in our SOM layer, which will make our network able to expand and contract if necessary. Also, we will extend the visualization mechanisms to the sound channels in a way that we can create a mechanism to hear what the network learned. Finally, we intend to further develop the network in HRI scenarios, where it will be used for giving the robot a deeper understanding of the emotional behavior of humans.

Acknowledgements

The authors would like to thank Katja Koesters for her constructive comments and insightful suggestions that improved the quality of this manuscript.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially supported by CAPES Brazilian Federal Agency for the Support and Evaluation of Graduate Education (p.n.5951–13–5), the German Research Foundation DFG under project CML (TRR 169), and the Hamburg Landesforschungsförderungsprojekt CROSS.

References

Abdel-Hamid, O., Mohamed, A.-R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*, 1533–1545.

Adolphs, R. (2002). Neural systems for recognizing emotion. *Current Opinion in Neurobiology*, *12*, 169–177.

Afzal, S., & Robinson, P. (2009). Natural affect data—Collection & annotation in a learning context. In *3rd international conference on affective computing and intelligent interaction* (pp. 1–7). Piscataway, NJ: IEEE Press. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5349537 (accessed 19 August 2016).

Banda, N., & Robinson, P. (2011). Noise analysis in audio-visual emotion recognition. In *13th international conference on multimodal interaction (ICMI '11)* (pp. 1–4). New York: ACM Press. Available at: http://citeseerx.ist.psu.edu/

viewdoc/summary?doi=10.1.1.228.6522 (accessed 19 August 2016).

Barros, P., Jirak, D., Weber, C., & Wermter, S. (2015a). Multimodal emotional state recognition using sequence-dependent deep hierarchical features. *Neural Networks*, *72*, 140–151.

Barros, P., Weber, C., & Wermter, S. (2015b). Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction. In *15th IEEE-RAS international conference on humanoid robots* (pp. 646–651). Piscataway, NJ: IEEE Press. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7363421 (accessed 19 August 2016).

Cabanac, M. (2002). What is emotion? *Behavioural Processes*, *60*, 69–83.

Castellano, G., Kessous, L., & Caridakis, G. (2008). Emotion recognition through multiple modalities: Face, body gesture, speech. In C. Peter & R. Beale (Eds.), *Affect and emotion in human-computer interaction* (pp. 92–103). Berlin, Germany: Springer.

Chen, S., Tian, Y., Liu, Q., & Metaxas, D. N. (2013). Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. *Image and Vision Computing*, *31*, 175–185.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, *18*, 32–80.

Dhall, A., Goecke, R., Joshi, J., Sikka, K., & Gedeon, T. (2014). Emotion recognition in the wild challenge 2014: Baseline, data and protocol. *16th international conference on multimodal interaction (ICMI '14)* (pp. 461–466). New York: ACM Press. Available at: http://dl.acm.org/citation.cfm?id=2666275 (accessed 19 August 2016).

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, *19*, 34–41.

Ekman, P. (2007). *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life*. Macmillan. Available at: http://psycnet.apa.org/psycinfo/2003-88051-000 (accessed 19 August 2016).

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*, 124–129.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., & Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, *11*, 625–660.

Essen, D. C. V., & Gallant, J. L. (1994). Neural mechanisms of form and motion processing in the primate visual system. *Neuron*, *13*, 1–10.

Foroni, F., & Semin, G. R. (2009). Language that puts you in touch with your bodily feelings: The multimodal responsiveness of affective expressions. *Psychological Science*, *20*, 974–980.

Fregnac, Y., Monier, C., Chavane, F., Baudot, P., & Graham, L. (2003). Shunting inhibition, a silent step in visual cortical computation. *Journal of Physiology*, 97(4), 441–451.

Friesen, E., & Ekman, P. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.

George, J., & Shamir, L. (2015). Unsupervised analysis of similarities between musicians and musical genres using spectrograms. *Artificial Intelligence Research*, *4*, 61–71.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *14th international conference on artificial intelligence and statistics (AISTATS-11)* (Vol. 15, pp. 315–323). Available at: http://www.jmlr.org/proceedings/papers/v15/glorot11a.html (accessed 19 August 2016).

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*, 20–25.

Grossberg, S. (1992). *Neural networks and natural intelligence*. Cambridge, MA: MIT Press.

Gunes, H., & Piccardi, M. (2006). A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *18th international conference on pattern recognition (ICPR)* (Vol. 1, pp. 1148–1153). Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1699093 (accessed 19 August 2016).

Gunes, H., & Piccardi, M. (2009). Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *39*, 64–84.

Hamlin, J. K. (2013). Moral judgment and action in preverbal infants and toddlers evidence for an innate moral core. *Current Directions in Psychological Science*, *22*, 186–193.

Haq, S., & Jackson, P. (2010). Multimodal emotion recognition. In W. Wang (Ed.), *Machine audition: Principles, algorithms and systems* (pp. 398–423). Hershey, PA: IGI Global.

Haq, S., Jackson, P. J., & Edge, J. (2009). Speaker-dependent audio-visual emotion recognition. In *2009 international conference on audio-visual speech processing (AVSP)* (pp. 53–58). Available at: https://scholar.google.de/scholar?cluster=5579645476741846741&hl=de&as_sdt=0,5 (accessed 19 August 2016).

Harter, S., & Buddin, B. J. (1987). Children's understanding of the simultaneity of two emotions: A five-stage developmental acquisition sequence. *Developmental Psychology*, *23*, 388–399.

Hau, D., & Chen, K. (2011). Exploring hierarchical speech representations with a deep convolutional neural network. In *11th UK workshop on computational intelligence (UKCI'11)* (p. 37). Available at: https://scholar.google.de/scholar?cluster=18130383993448916657&hl=de&as_sdt=0,5 (accessed 19 August 2016).

Hickok, G. (2012). The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. *Journal of Communication Disorders*, *45*, 393–402.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurons in the cat's striate cortex. *Journal of Physiology*, *148*, 574–591.

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D Convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 221–231.

Jin, Q., Li, C., Chen, S., & Wu, H. (2015). Speech emotion recognition with acoustic and lexical features. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4749–4753). Piscataway,

NJ: IEEE Press. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7178872 (accessed 19 August 2016).

Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, C., Memisevic, R., … Wu, Z. (2013). Combining modality specific deep neural networks for emotion recognition in video. In *15th international conference on multimodal interaction (ICMI '13)* (pp. 543–550). New York: ACM Press. Available at: http://dl.acm.org/citation.cfm?id=2531745 (accessed 19 August 2016).

Karnowski, T. P., Arel, I., & Rose, D. (2010). Deep spatio-temporal feature learning with application to image classification. In *9th international conference on machine learning and applications (ICMLA)* (pp. 883–888). Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5708961 (accessed 19 August 2016).

Khalil-Hani, M., & Sung, L. S. (2014). A convolutional neural network approach for face verification. In *2014 international conference on high performance computing simulation (HPCS)* (pp. 707–714). Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6903759 (accessed 19 August 2016).

Kim, S. M., Valitutti, A., & Calvo, R. A. (2010). Evaluation of unsupervised emotion models to textual affect recognition. In *NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 62–70). Association for Computational Linguistics.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*, 1464–1480. Available at: http://dl.acm.org/citation.cfm?id=1860639 (accessed 19 August 2016).

Kret, M. E., Roelofs, K., Stekelenburg, J. J., & de Gelder, B. (2013). Emotional signals from faces, bodies and scenes influence observers' face expressions, fixations and pupil-size. *Frontiers in Human Neuroscience*, *7*, 810. Available at: mhttp://journal.frontiersin.org/article/10.3389/fnhum.2013.00810/full (accessed 19 August 2016).

Lawrence, S., Giles, C., Tsoi, A. C., & Back, A. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, *8*, 98–113.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=726791 (accessed 19 August 2016).

Lewis, M. (2012). *Children's emotions and moods: Developmental theory and measurement*. Springer. Available at: https://scholar.google.de/scholar?q=Children%E2%80%99s+emotions+and+moods%3A+Develop-+mental+theory+and+measurement&btnG=&hl=de&as_sdt=0%2C5 (accessed 19 August 2016).

Li, T. L., Chan, A. B., & Chun, A. H. (2010). Automatic musical pattern feature extraction using convolutional neural network. In *International multiconference of engineers and computer scientists (IMECS 2010)* (Vol. 1). Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.302.7795 (accessed 19 August 2016).

Liu, M., Chen, H., Li, Y., & Zhang, F. (2015). Emotional tone-based audio continuous emotion recognition. In X. He, S. Luo, D. Tao, C. Xu, J. Yang, & M. Hasan (Eds.), *Multimedia modeling* (pp. 470–480). Springer. Available at: http://link.springer.com/chapter/10.1007/978-3-319-14442-9_52 (accessed 19 August 2016).

Liu, M., Wang, R., Li, S., Shan, S., Huang, Z., & Chen, X. (2014). Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild. In *16th international conference on multimodal interaction* (pp. 494–501). New York: ACM Press. Available at: http://dl.acm.org/citation.cfm?id=2666274 (accessed 19 August 2016).

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *5th Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). Oakland, CA. Available at: https://scholar.google.de/scholar?start=0&hl=de&as_sdt=0,5&cluster=14924728719521477429 (accessed 19 August 2016).

Pons, F., Harris, P. L., & de Rosnay, M. (2004). Emotion comprehension between 3 and 11 years: Developmental periods and hierarchical organization. *European Journal of Developmental Psychology*, *1*, 127–152.

Ringeval, F., Amiriparian, S., Eyben, F., Scherer, K., & Schuller, B. (2014). Emotion recognition in the wild: Incorporating voice and lip activity in multimodal decision-level fusion. In *16th international conference on multimodal interaction (ICMI '14)* (pp. 473–480). New York: ACM Press. Available at: http://dl.acm.org/citation.cfm?id=2666271 (accessed 19 August 2016).

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*, 145–172.

Sainath, T. N., Kingsbury, B., Saon, G., Soltau, H., Rahman Mohamed, A., Dahl, G., & Ramabhadran, B. (2015). Deep convolutional neural networks for large-scale speech tasks. *Neural Networks*, *64*, 39–48.

Sariyanidi, E., Gunes, H., & Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*, 1113–1133.

Schluter, J., & Bock, S. (2014). Improved musical onset detection with convolutional neural networks. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP 2014)* (pp. 6979–6983). Piscataway, NJ: IEEE Press. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6854953 (accessed 19 August 2016).

Sloman, A. (2001). Beyond shallow models of emotion. *Cognitive Processing*, *2*, 177–198.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*, 293–302.

Ultsch, A. (2003). U-Matrix: A tool to visualize clusters in high dimensional data. Report, University of Marburg, Germany, December.

Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent Data Analysis*, *3*, 111–126.

Viola, P., & Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, *57*, 137–154.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *13th european conference of computer vision (ECCV 2014)* (pp. 818–833). Berlin, Germany: Springer. Available at: http://link.springer.com/chapter/10.1007/978-3-319-10590-1_53 (accessed 19 August 2016).

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*, 39–58.

About the Authors

Pablo Barros received his Bachelor degree in Information Systems from the Federal Rural University of Pernambuco and his Master degree in Computer Engineering from the University of Pernambuco, both in Brazil. Since 2013, he has been a research associate and PhD candidate in the Knowledge Technology Group at the University of Hamburg, Germany, where he is part of the research project CML (crossmodal learning). His main research interests include deep learning and neurocognitive systems for multimodal emotional perception and learning, human–robot interaction and cross-modal neural architectures.

Stefan Wermter received a Diplom from the University of Dortmund, an MSc from the University of Massachusetts, and a Doctorate and Habilitation from the University of Hamburg, all in Computer Science. He was a visiting research scientist at the International Computer Science Institute in Berkeley before leading the Chair in Intelligent Systems at the University of Sunderland, UK. Currently Stefan Wermter is Full Professor in Computer Science at the University of Hamburg and Director of the Knowledge Technology institute. His main research interests are in the fields of neural networks, hybrid systems, cognitive neuroscience, bio-inspired computing, cognitive robotics and natural language processing. In 2014 he was general chair for the International Conference on Artificial Neural Networks (ICANN). He is also on the current board of the European Neural Network Society, and associate editor of the journals *Transactions on Neural Networks and Learning Systems, Connection Science International Journal for Hybrid Intelligent Systems*, and *Knowledge and Information Systems*. He is on the editorial board of the journals *Cognitive Systems Research* and *Journal of Computational Intelligence*.