In: Proc. of the IEEE-RAS International Conference on Humanoids Robots (HUMANOIDS), Seoul, Korea (2015)

Emotional Expression Recognition with a Cross-Channel Convolutional Neural Network for Human-Robot Interaction

Pablo Barros, Cornelius Weber and Stefan Wermter

Abstract— **The study of emotions has attracted considerable attention in several areas, from artificial intelligence and psychology to neuroscience. The use of emotions in decision-making processes is an example of how multi-disciplinary they are. To be able to communicate better with humans, robots should use appropriate communicational gestures, considering the emotions of their human conversation partners. In this paper we propose a deep neural network model which is able to recognize spontaneous emotional expressions and to classify them as positive or negative. We evaluate our model in two experiments, one using benchmark datasets, and the other using an HRI scenario with a humanoid robotic head, which itself gives emotional feedback.**

I. Introduction

Emotion recognition has become one of the major topics in human-robot interaction in recent years. A robot with the capability of recognizing human emotions can behave and adapt to different communication situations [1], or take into account the emotions to induce a specific behavior in the robot as a special subsystem [2]. Besides communication, humans collect, analyze and distinguish emotional expressions of other humans as part of the decision making process [3]. A robot which could mimic this judgmental capability could be able to enhance its interactions skills, realize complex tasks and even create a certain discernment about the information it is receiving. Imagine that a robot can evaluate the quality of the information it is receiving from a person, and use this evaluation to make a decision [4], or is able to adapt its behavior based on emotion expressed by humans [5].

Positive and negative emotions have an important impact in human development. The broaden-and-build theory, proposed by Fredrickson et al. [6], states that positive emotions enhance the personal relation between a thought and a reaction, which tunes their personal, physical, intellectual, social and psychological skills. A similar development is the expression of negative emotions, which influences our self-awareness of the environment and situations [7]. A robot which is to recognize these expressions has a strong ability, from acting as a human avatar, a tool to solve tasks or even a human partner. When imitating human learning and showing emotional feedback, a robot can achieve an improved level of interaction skills being able to collaborate with a human in a social environment as a teammate [8].

The recognition of human emotions by robots is not a simple task. In a non-verbal communication, two or more modalities, such as facial expression and body posture, are complementary [9] and when presented together, both modalities are recognized differently than when they are shown individually. The observation of these modalities provides better accuracy in emotion perception [10], and is the focus of promising work in automatic emotion recognition systems [11]. In this work, Chen et al. are able to recognize emotions from different persons, but to do so they use several pre-processing, feature extraction and fusion techniques, which leads to high computational costs and makes the problem limited to each of the techniques' limitations. This usually makes the process unsuitable to be used in a real human-robot interaction scenario.

Most of the work on automatic emotional recognition does not distinguish between positive and negative emotions. Most of the research in this area is based on verbal communication, like the work of Pavaloi et al. [12], which extracts emotions from spoken vowels, and Tahon et al. [13], which uses a robot to detect emotions from human voices. Some other research introduces similar methods, but they are applied to basic emotion recognition, such as anger, happiness, sadness and several other emotions [14]. They evaluate their system using a controlled environment, and are able to identify the intensity of each emotion in an acted dataset, where each emotion is performed the same way by different subjects. The work of Ballihi et al. [15] detects positive/negative emotion expressions from RGB-D data. They use a method to classify the intensity of each expression using multimodal information, from the upper-body motion and face expression. Their method extracts different features from the face expression, and uses the depth information from the upper-body motion to create a velocity vector. These features are fed to a classifier based on random forest techniques. In their method, several feature extraction techniques are used, each one of them coding one aspect of the data: action units from the eyebrows and mouth and the distance of the person from the camera, to extract the upper-body motion. At the end, all these features are collected individually. Their model is also extremely dependent on the position of the person on the image, illumination, and image pre-processing to detect eyes and face regions.

To solve restrictions as illumination, pre-processing constraints, and position of the person, which are present in the mentioned models, our method implements a multimodal emotion recognition system based on spatial-temporal hierarchical features. Our neural architecture, named Cross-

The authors are with the University of Hamburg - Department of Computer Science, Vogt-Koelln-Strasse 30, 22527 Hamburg - Germany.
barros, weber, wermter
@informatik.uni-hamburg.de

Channel Convolutional Neural Network (CCCNN), extends the power of Convolutional Neural Networks (CNN) [16] to multimodal data extraction and is able to learn and extract general and specific features of emotions based on face expression and body motion. The first layers of our network extract low-level features, such as edges and motion directions. These features are passed to deeper layers, which build a cross-modal representation of the data based on cross-convolution channels. This characteristic allows our network to learn the most influential features of each modality, instead of using specific feature extractors pre-built into the system.

We evaluate our model in two experiments. The first one evaluates the network on benchmark datasets. We use two categories of datasets: one with acted emotions, and one with spontaneous emotions. In the second experiment, the network is deployed on a humanoid robot head and applied in a human-robot interaction (HRI) scenario, where the robot has to classify the emotions expressed by a subject. The features that the network learns, the behavior of the model when different data is presented, and the robustness of the model when applied in a human-robot interaction scenario are evaluated and discussed.

The paper is structured as follows: The next section introduces our neural model and describes how temporal and spatial features are learned and extracted. The methodology for our experiments, the results and discussion are given in Section III. Finally, the conclusion and future work are given in Section IV.

II. Proposed model

Our neural network is based on Convolutional Neural Networks (CNN) to extract and learn hierarchical features. A CNN simulates the simple and complex cells present in the visual cortex of the brain [17], [18] by applying two operations: convolution and pooling. The simple cells, represented by the convolution operations, convolve the image using local filters to compute high-order features. These features are learned, and are able to extract different features depending on where they are applied. The complex cells generate scale invariance by pooling simple cell activations into a new smaller image grid. The activation of each simple cell v_{nc}^{xy} at the position (x,y) of the n^{th} receptive field in the c^{th} layer is given by

$$v_{nc}^{xy} = max\left(b_{nc} + \sum_{m}\sum_{h=1}^{H}\sum_{w=1}^{W} w_{(c-1)m}^{hw} v_{(c-1)m}^{(x+h)(y+w)}, 0\right), \tag{1}$$

where $max(\cdot, 0)$ implements the rectified linear function, which was shown to be more suitable than other functions for training deep neural architectures [19], b_{nc} is the bias for the nth feature map of the cth layer,and m indexes over the set of feature maps in the $c-1$ layer connected to the current layer c. In equation (1), $w_{(c-1)m}^{hw}$ is the weight of the connection between each unit within a receptive field, defined by (h,w), connected to the previous layer, $c-1$, and to the filter map m. H and W are the height and width of the receptive field.

In the complex cell layer, a receptive field of the previous simple cell layer is connected to a unit in the current layer, which reduces the dimensionality of the feature maps. For each complex cell layer, only the maximum value of non-overlapping patches of the input feature map are passed to the next layer. This enhances invariance to scale and distortions of the input, as described in [20].

A. Temporal features

To be able to extract temporal features, based on the image change through time, we apply a cubic receptive field [21] to a sequence of images. The cubic receptive field applies complex cells to the same region of a stream of visual stimuli. This process extracts the changes within the sequence, coding the temporal features in a series of different representations. In a cubic convolution, the value of each simple cell (x,y,z) at the n^{th} receptive field in the c^{th} layer is defined as:

$$\begin{aligned} v_{nc}^{xyz} =& max(b_{nc}+ \\ & \sum_{m}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{r=1}^{R} w_{(c-1)m}^{hwr} v_{(m-1)}^{(x+h)(y+w)(z+r)}, 0), \end{aligned} \tag{2}$$

where $max(\cdot, 0)$ represents the rectified linear function, b_{cn} is the bias for the nth filter map of the cth layer, and m indexes the set of feature maps in the (c-1) layer connected to the current layer c. In equation (2), $w_{(c-1)m}^{hwr}$ is the weight of the connection between the unit (h,w,r) within a receptive field connected to the previous layer ($c-1$) and the filter map m. H and W are the height and width of the receptive field, z indexes each image in the image stack, and R is the number of consecutive image frames stacked together representing the time dimension of the receptive field.

B. Inhibitory receptive fields

A problem shared among deep neural network architectures is the large amount of computational power used for training. Usually several layers of simple and complex cells are necessary to learn general feature representations, which increases the number of parameters to be updated during training. To reduce the number of layers in the network, we introduce the use of shunting inhibitory fields [22] in deeper layers. Shunting inhibitory neurons are neuro-physiological plausible mechanisms that are present for several visual and cognitive functions [23]. When applied in complex cell structures on a CNN [24], shunting neurons can derive filters that are more robust to geometric distortions. Each shunting neuron S_{nc}^{xy} at the position (x,y) of the n^{th} receptive field in the c^{th} layer is activated as:

$$S_{nc}^{xy} = \frac{v_{nc}^{xy}}{a_{nc} + V_{nc}^{xy}} \tag{3}$$

where v_{nc}^{xy} is the activation of the unit in the position of the receptive field and V_{nc}^{xy} is the activation of the inhibitory neuron. Note that each inhibitory neuron has its own set of weights, which are also trained with backpropagation. The passive decay term is defined as a_{nc}, is a defined parameter and is the same for all shunting neurons in the model.

The shunting neurons have more complex decision boundaries and thus specify the filter tuning. When applied to the first simple layers, we found that these neurons create very specific edge-detectors which will not be able to extract general features, but when applied to deeper layers they create a stronger filter for shapes. We understand strong filters as filters that are robust enough to extract information from images with different backgrounds, different subjects and different camera positions.

C. Cross-Channel Convolutional Neural Network (CCCNN)

To be able to deal with multimodal data, our network uses the concept of the Multichannel Convolutional Neural Networks (MCCNN) [25]. In the MCCNN architecture, several channels, each one of them composed by a different CNN, are connected at the end to a hidden layer, and trained as one single architecture. Inspired by the primate visual cortex model described by Van Essen et al. [26], our network has two channels. The first channel is responsible for learning and extracting information based on the contour, shape and texture of a face, which mimics the encoding of information in the V1 area of the primate visual cortex. The second channel codes information about the orientation, direction and speed of changes within the faces of a sequence, similar as the information coded by the V2 area. Figure 1 illustrates the final Cross Channel Convolutional Neural Network (CCCNN) topology, using different channels.

The first channel implements a common convolutional neural network with two layers. The first layer is composed of 10 receptive fields, each of them connected with a max-pooling layer. Each receptive field has a size of 5x5, and the pooling layer reduces the dimensionality of the data by a factor of 2. The second layer contains 20 receptive fields, each of them with a size of 7x7, and connected to a pooling layer that reduces the data dimensionality by a factor of 2. The second layer implements inhibitory receptive fields, which have the same topology. We feed this channel with images of the face, one at a time. To extract the face of the images, we use an Adaboost-based face detector [27], works in real-time and is robust to illumination and translation.

The second channel is composed of one layer with a cubic receptive field implementation. This layer is able to extract spatial-temporal features from a sequence of images. It has 10 cubic receptive fields, and each one has a size of 5x5x4, which means that we feed this channel with 4 images per time step.

A similar approach was evaluated for large-scale video classification, when different channels received the same image, but in different resolutions [28]. In this work, each channel received the same image, being able to use the multichannel architecture to extract different representation from the same input. We differ by introducing the concept of multi-channel learning. Our cross-channel encodes similar information as the V4 area discussed in the Van Essen et al. [26] model. This channel receives as input the features coming from channel 1 and channel 2 and processes them with a layer of simple and complex cells. The filters learned by the cross-channel are able to code the most important features for emotion recognition: face expression shapes through time correlated with body motion. Different than manually fusing motion information with face expression, this layer is able to learn how to derive specific patterns for emotion recognition. Our cross-channel has 10 receptive fields, each one with the size of 5x5. Each of them is connected with a max-pooling layer which reduces the dimensionality by a factor of 2.

Because our model applies topological convolution in the images, the choice of the size of the receptive field has an important impact in the learning process. The receptive fields in our cross-channel should be large enough to be able to capture the whole concept of the image, and not only part of it. With a small receptive field, our cross learning will not be able to capture the concept of the face and the motion, and only correlate some regions of the face image with the same regions in the sequence images. In our experiments, each output neuron of the first channel is related to an area of 32x32 pixels from the original image and in the second channel to 15x15. This means, that each receptive field of our cross-channel is applied to a region of 36x36 pixels from the face image, and 19x19 from the sequence image, reaching more than half of the image per receptive field.

D. Emotion expression classification

To classify the features extracted by our convolutional network-based layers, our cross-channel is connected to a fully connected hidden layer. The hidden layer has 250 units, with a rectified linear activation function. The hidden layer is connected to an output layer which implements a softmax classifier. The softmax function is a generalization of the logistic function which represents the probability that the input images belong to a certain class.

Based on experimental results, we found that using a sequence of 4 frames as input produced optimal results. Each image has a size of 64x48 pixels. The first channel receives only one image, with the face of the subject extracted from the second frame of the sequence, and the second channel receives the whole sequence. This way, the face processed by the first channel is part of the sequence seen by the second channel. The parameters were chosen based on the best experimental results. To be able to classify the probability of each emotion expression, our network categorizes each sequence into 3 outputs: Positive, Negative and Neutral.

To improve the filter tuning, we applied some regularization techniques during training such as applying a dropout technique [29], which shuts down random units during training and makes the learning process more robust in our hidden layer. We also apply L1 regularization and use a momentum term during the training, which helped to decrease training error.

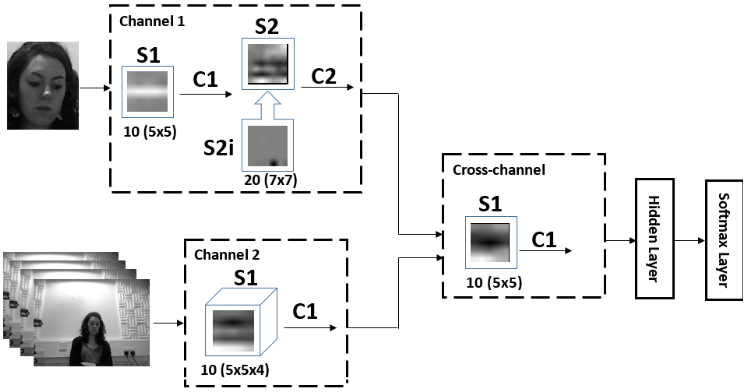

Fig. 1. Illustration of the proposed cross-channel convolutional neural network, with examples of filters after training the network. S represents the simple cells, C represents the complex cells. The first channel receives as input a face and its second layer implements a shunting inhibition receptive field, represented as S2i. Channel 2 receives a sequence and implements a cubic receptive field. The cross-channel receives the information from both channels, and is the input to a fully connected hidden layer, which is connected to the final softmax classifier layer.

III. Experiments

A. Methodology

To evaluate our model we propose two experimental setups: benchmark trials and a HRI scenario.

1) Benchmark trials: The benchmark trials are to evaluate the filters of the network by training them with two different datasets, one with acted emotional expressions, and the other showing spontaneous emotions. The results are collected and the behavior of the network is evaluated. The first dataset we used was the Cohn-Kanade dataset [30]. This corpus contains 7 expressions of emotions, performed by 123 different subjects. They are labeled as *Angry*, *Contemptuous*, *Disgusted*, *Feared*, *Happy*, *Sad* and *Surprised*. Each example of emotion contains a sequence with 10 to 60 frames, and starts in the onset (neutral face) and continues until the peak of the facial expression (the offset). Figure 2 (a) shows examples of frames of a sequence in the Cohn-Kanade dataset.

The second dataset is the 3D corpus of spontaneous complex mental states (CAM3D) [31]. The corpus is composed of 108 video/audio recordings from 7 subjects and in different indoor environments. Each video exhibits the upper body of one subject while the emotion expression is performed. Each subject demonstrates the emotions in a natural and spontaneous way, without following any previously shown pose. The corpus contains a total of 12 emotional states, which were labeled using crowd-sourcing: *Agreeing*, *Bored*, *Disagreeing*, *Disgusted*, *Excited*, *Happy*, *Interested*, *Neutral*, *Sad*, *Surprised*, *Thinking* and *Unsure*. Figure 2 (b) shows an example of a sequence in the CAM3D dataset.

To give our model the capability to identify the emotion expression as negative or positive, we separated both datasets using the emotion annotation and representation language (EARL) [32], which classifies 48 emotions into negative and positive expressions. Table I shows how each emotion in each dataset was classified. Besides positive and negative, we created a neutral category which indicates if the person is exhibiting no emotion at all. This category gives the model the capability to identify if the person is exhibiting an emotion or not. For the Cohn-Kanade dataset we use the four first frames of each emotion, and label them as neutral. For the CAM3D dataset, we use the already given neutral label.

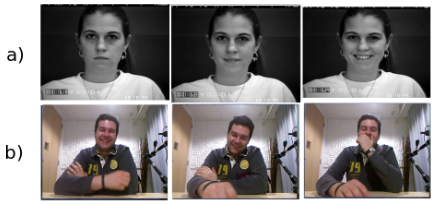

Fig. 2. Examples of sequences: (a) Cohn-Kanade (b) CAM3D datasets.

To evaluate the robustness of the features learned by our network, we executed four experiments. In the first two experiments, the network was trained and tested with each of the datasets separated into 60% of the data for training, and 40% for testing. The second set of experiments used one of the datasets to train the network and the other to test it. In this case, all the data from one dataset was used for training, and all the data of the other dataset for testing. We ran each experiment 30 times and collected accuracy and standard deviation.

2) HRI Scenario: To evaluate our model in a real HRI scenario, we trained the network with both benchmark datasets and deployed it in a humanoid robotic head. We used the head of an iCub robot [33], which has a common

TABLE I

Separation of the emotions in the Cohn-Kanade and CAM3D datasets based on EARL[32].

Dataset	Positive	Negative
Cohn-Kanade[30]	*Happy* and *Surprised*	*Angered*, *Disgusted*, *Contemptuous*, *Feared* and *Sad*
CAM3D [31]	*Agreeing*, *Excited*, *Happy*, *Interested*, *Surprised* and *Thinking*	*Bored*, *Disagreeing*, *Disgusted*, *Sad* and *Unsure*

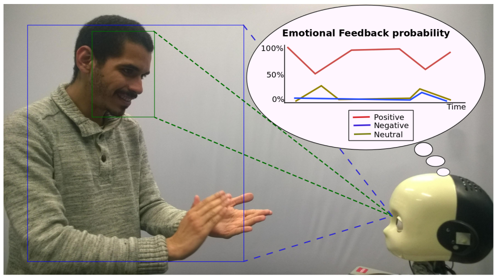

Fig. 3. Our HRI experiment scenario. A person was in front of the iCub head and performed an emotional expression. The robot identified the emotion that was displayed and gave a proper feedback: a smile, an angry or a neutral face. The green square indicates the face image, used as input for the first channel of the CCCNN, and the blue square indicates the sequence region, used as input for the second channel.

RGB camera. A subject was in front of the robot and presented one emotional state (positive, negative or neutral). The robot recognized the emotional state and gave feedback by changing its mouth and eyebrow LEDs, indicating a smile for positive, an angry face for negative, or a neutral face, when there was no emotion expressed.

In this experiment, we used 5 different subjects. Each subject performed 10 different emotional expressions in front of the robot. The subject was instructed to perform a different emotion per time (positive, negative or neutral), but had no indication of how it had to perform it. This way, every expression was spontaneous and not structured. Figure 3 shows an illustration of the scenario.

B. Results

We collected the mean accuracy for all experiments of the benchmark trials. When using the CAM3D dataset, we repeated the same experiments present in [15]. In their fusion of motion and face expression, they obtained a mean accuracy of 71.3%, while our model was able to achieve 86.2%. When training and testing the model with the Cohn-Kanade dataset, the mean accuracy was 92.5%.

Training the model with the Cohn-Kanade dataset and testing with the CAM3D showed the lowest accuracy rate, a total of 70.5%. The other way, using the CAM3D dataset for training and the Cohn-Kanade for testing, we obtain a mean accuracy of 79.5%. Table II shows our results.

When applied to the HRI scenario, we collect the mean accuracy of all 5 subjects, each one performing 10 emotional expressions. Table III shows the mean accuracy and the standard deviation for all the subjects. A total mean accuracy of 74.2 % was obtained. A computer with an Intel XEON CPU processor with 2.4Ghz was used to recognize the images from the iCub and took 0.42ms in average to recognize each expression.

IV. Analysis and Conclusions

The use of emotional expression recognition improves how robots can communicate and react to humans in several scenarios. Especially the recognition of positive/negative emotions can be applied to decision-making tasks. For example, it is possible to identify when a human is giving emotional information. Giving a robot this ability can improve how intelligent agents analyze information. But to do so, a strong emotion recognition system should be implemented. This paper proposes a system that can recognize spontaneous emotions from different subjects and can be used in human-robot interaction scenarios.

Our model is based on a Convolutional Neural Network (CNN) which uses its capability to learn which are the most influential features for emotion expression recognition. Our model can deal with two processes at the same time: extract features from the contour, shape and face characteristics and gather information about the motion of a subject. We introduce the use of shunting inhibitory receptive fields to increase the capacity of deeper levels of the network to extract strong and general features, and the use of cross-channel learning to create correlations between the static and dynamic streams.

We evaluate our model in two different experiments: First, we use two kinds of datasets to train the network: The Cohn-Kanade dataset which contains examples of acted emotions and the CAM3D corpus which contains spontaneous examples. We can observe that our model is able to recognize both acted and spontaneous emotion expressions. In the second experiment, we deploy our network on a humanoid robotic head which is able to identify positive/negative emotion expressions from a subject. In this scenario, the robotic head is able to react to each recognized emotional expression and to give an emotional feedback. The network is able to recognize emotions from different environments, different subjects performing spontaneous expressions, and in real-time, which are basic characteristics for real human-robot interaction scenarios.

One of the attractive properties of our network is the capability to learn which characteristics of an emotional expression are the most influential for the positive/negative emotion classification. As shown in the experiments of the benchmark trial, our network was able to identify both acted and spontaneous emotions. When the datasets are mixed for training and testing, the network was able to identify emotional characteristics even in spontaneous examples. Our

TABLE II

Accuracy and Standard deviation for the benchmark trial experiments.

Test Train	Cohn-Kanade	CAM3D
Cohn-Kanade	92.5% (+/- 2.5)	70.5% (+/- 3.3)
CAM3D	79.5% (+/- 3.1)	86.29% (+/- 1.8)

TABLE III

Accuracy and Standard deviation for the HRI scenario experiments.

Subject	1	2	3	4	5	Total
Acc	70.2%	75.7%	73.9%	78.7%	72.5%	74.2%
Std	3.8	2.7	3.2	2.9	3.1	2.8

experiments show that spontaneous emotions have some sort of structure, even if executed by different persons, since the network is able to classify spontaneous emotion expressions learned from acted examples. This was even clearer in our HRI scenario, were the network was able to identify the emotions expressed by subjects that were not present in the dataset, and in a spontaneous way.

For future work, we plan to further develop our network in a decision making scenario, where the recognized emotions could be used to weight the veracity of the received information. We also plan to extend the model to work with audio inputs together with visual stimuli.

ACKNOWLEDGMENT

This work was partially supported by the CAPES - Brazilian Federal Agency for the Support and Evaluation of Graduate Education (p.n.5951–13–5). The authors thank Dr. Peter Robinson for providing the CAM3D Corpus.

REFERENCES

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

[2] S. Tokuno, G. Tsumatori, S. Shono, E. Takei, G. Suzuki, T. Yamamoto, S. Mituyoshi, and M. Shimura, "Usage of emotion recognition in military health care," in *Defense Science Research Conference and Expo (DSR), 2011*, Aug 2011, pp. 1–5.

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

[4] J. S. Lerner and D. Keltner, "Beyond valence: Toward a model of emotion-specific influences on judgement and choice," *Cognition and Emotion*, vol. 14, no. 4, pp. 473–493, 2000.

[5] B. Schuller, G. Rigoll, S. Can, and H. Feussner, "Emotion sensitive speech control for human-robot interaction in minimal invasive surgery," in *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*, Aug 2008, pp. 453–458.

[6] B. L. Fredrickson, "The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions." *American psychologist*, vol. 56, no. 3, p. 218, 2001.

[7] N. D. Cook, *Tone of Voice and Mind: The Connections Between Intonation, Emotion, Cognition and Consciousness*. John Benjamins, 2002.

[8] C. Breazeal and R. Brooks, "Robot emotions: A functional perspective," in *Who Needs Emotions*, J. Fellous, Ed. Oxford University Press, 2004.

[9] Y. Gu, X. Mai, and Y.-j. Luo, "Do bodily expressions compete with facial expressions? time course of integration of emotional signals from the face and the body," *PLoS ONE*, vol. 8, no. 7, pp. 62–67, 07 2013.

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

[11] S. Chen, Y. Tian, Q. Liu, and D. N. Metaxas, "Recognizing expressions from face and body gesture by temporal normalized motion and appearance features," *Image and Vision Computing*, vol. 31, no. 2, pp. 175 – 185, 2013.

[12] I. Pavaloi, A. Ciobanu, M. Luca, E. Musca, T. Barbu, and A. Ignat, "A study on automatic recognition of positive and negative emotions in speech," in *System Theory, Control and Computing (ICSTCC), 2014 18th International Conference*, Oct 2014, pp. 221–224.

[13] M. Tahon, A. Delaborde, and L. Devillers, "Real-life emotion detection from speech in human-robot interaction: Experiments across diverse corpora with child and adult voices." in *INTERSPEECH*. ISCA, 2011, pp. 3121–3124.

[14] D. A. Gómez Jáuregui and J.-C. Martin, "Evaluation of vision-based real-time measures for emotions discrimination under uncontrolled conditions," in *Proceedings of the 2013 on Emotion Recognition in the Wild Challenge and Workshop*, ser. EmotiW '13. New York, NY, USA: ACM, 2013, pp. 17–22.

[15] L. Ballihi, A. Lablack, B. Amor, I. Bilasco, and M. Daoudi, "Positive/negative emotion detection from RGB-D upper body images," in *Face and Facial Expression Recognition from Real World Videos*, ser. Lecture Notes in Computer Science, Q. Ji, T. B. Moeslund, G. Hua, and K. Nasrollahi, Eds. Springer International Publishing, 2015, vol. 8912, pp. 109–120.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324.

[17] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurons in the cat's striate cortex," *Journal of Physiology*, vol. 148, pp. 574–591, 1959.

[18] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition," *Neural Networks*, vol. 1, no. 2, pp. 119 – 130, 1988.

[19] X. Glorot, A. Bordes, and Y. Bengio, "Deep Sparse Rectifier Neural Networks," 2011, pp. 315–323.

[20] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *In proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012*, 2012, pp. 3642–3649.

[21] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 221–231, Jan 2013.

[22] Y. Fregnac, C. Monier, F. Chavane, P. Baudot, and L. Graham, "Shunting inhibition, a silent step in visual cortical computation," *Journal of Physiology*, pp. 441–451, 2003.

[23] S. Grossberg, *Neural Networks and Natural Intelligence*. Cambridge, MA, USA: MIT Press, 1992.

[24] F. H. C. Tivive and A. Bouzerdoum, "A shunting inhibitory convolutional neural network for gender classification," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 4, 2006, pp. 421–424.

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

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

[27] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

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

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[30] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, June 2010, pp. 94–101.

[31] M. Mahmoud, T. Baltruaitis, P. Robinson, and L. Riek, "3d corpus of spontaneous complex mental states," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Springer Berlin Heidelberg, 2011, vol. 6974, pp. 205–214.

[32] M. Schrder, H. Pirker, M. Lamolle, F. Burkhardt, C. Peter, and E. Zovato, "Representing emotions and related states in technological systems," in *Emotion-Oriented Systems - The Humaine Handbook*, P. Petta, R. Cowie, and C. Pelachaud, Eds. Springer, 2011, pp. 367–386.

[33] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The icub humanoid robot: An open platform for research in embodied cognition," in *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, ser. PerMIS '08. New York, NY, USA: ACM, 2008, pp. 50–56.