2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Congress Center Hamburg
Sept 28 - Oct 2, 2015. Hamburg, Germany

Recognizing Complex Mental States With Deep Hierarchical Features For Human-Robot Interaction

Pablo Barros[1] and Stefan Wermter[1]

Abstract— **The use of emotional states for Human-Robot Interaction (HRI) has attracted considerable attention in recent years. One of the most challenging tasks is to recognize the spontaneous expression of emotions, especially in an HRI scenario. Every person has a different way to express emotions, and this is aggravated by the complexity of interaction with different subjects, multimodal information and different environments. We propose a deep neural model which is able to deal with these characteristics and which is applied in recognition of complex mental states. Our system is able to learn and extract deep spatial and temporal features and to use them to classify emotions in sequences. To evaluate the system, the CAM3D corpus is used. This corpus is composed of videos recorded from different subjects and in different indoor environments. Each video contains the recording of the upper-body part of the subject expressing one of twelve complex mental states. Our system is able to recognize spontaneous complex mental states from different subjects and can be used in such an HRI scenario.**

I. INTRODUCTION

Recognition of emotional states has become an important topic for human-robot interaction in recent years. By determining emotional states, robots can extend the ways of communication with humans, being able to approximate human-human communication, identify human behavior or extend interaction possibilities. Emotion-sensitive robots can be aware of how humans behave and adapt to the situation [1] and use emotion perception to act as specialist systems [2]. When a robot is able to perceive and react to emotions, human interaction also changes. Spexard et al. [3] discuss how humans react when interacting with an anthropomorphic robot and how their reactions change when the robot recognizes and expresses emotions. They conclude that when emotions are expressed, humans are more confident and act naturally which improves the success in their human-robot interaction scenario.

As discussed by [4] there is no consensus in the literature to define emotions, but the observation of several characteristics and among them facial expressions, are used to identify them. That explains why most of the applications using emotions for Human-Robot Interaction (HRI) use the facial basic emotions [5]: "disgust", "fear", "happiness", "surprise", "sadness" and "anger".

Although these emotions are described as universal and present in many forms of interaction, humans usually express themselves using the combination of one or more emotions, which represent complex mental states such as attitudes, cognitive states and intentions. Among these complex mental states are the expressions of neutral states [6] and more diverse emotions such as confusion, surprise and concentration [7]. These complex emotions extend the universal concepts described by Ekman et al. [5] and the capability of understanding them improves the way we can use emotions in HRI scenarios. For example, to perceive sarcasm in a dialogue one could change the reaction of an attendance robot.

In order to recognize emotions it is necessary to understand spontaneous behavior. Expressing emotions spontaneously, the subject can act naturally and express them in different ways, especially when non-verbal communication is used [8]. However, perceiving emotions by spontaneous expressions is a challenging task, and most of the automatic face recognition systems proposed in the literature cannot deal with it. For non-verbal communication, slight changes of body posture and face expressions can lead to completely different emotions.

Non-verbal interaction is a challenging part of HRI due to environmental noise, technical restrictions or the natural way to express and perceive commands or dialogues. For non-verbal emotion perception, the presence of facial expressions and body posture, especially upper-body motion, are complementary [9]. The observation of both modalities could provide a better accuracy in emotion perception [10]. However, there are only few approaches in the literature [11], [12] that deal with multimodal non-verbal emotion recognition, but none of them can deal with spontaneous emotions.

The human brain is capable of correlating information from different areas and thus recognizing emotions using different streams of information [13]. Facial transformations, past experiences and motion perception are used to generate a representation of the visual stimuli. Processing this information in computer systems was achieved by neural models [14], particularly ones which are able to create a hierarchy of feature representations such as Convolutional Neural Networks [15].

Convolutional Neural Networks (CNN) [16] are inspired by the hierarchical process of simple and complex cells in the human brain which extract and infer different information from visual stimuli. Each layer of the CNN can extract unique information from the stimuli and when stacked together these layers generate a complex representation of the visual stimuli. The first layers of a CNN act as edge detectors which are able to enhance simple characteristics such as border and pattern contrast. Deep layers receive the

[1]Pablo Barros and Stefan Wermter are with University of Hamburg, Department of Informatics, Vogt-Kolln-Strae 30, 22527 Hamburg, Germany {barros, wermter}@informatik.uni-hamburg.de

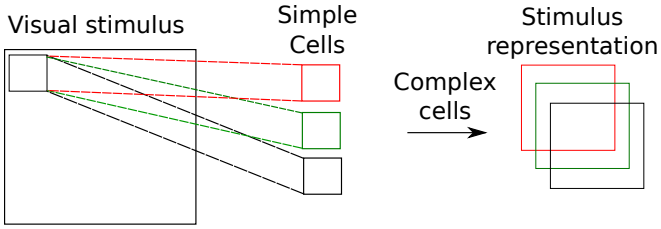

Fig. 1. Each layer of the CNN is composed of simple and complex cells, simulated using convolution and pooling operations. There are different filters representing simple cells that, when applied to the visual stimulus, generate a different representation. The pooling process reduces the dimensionality of the stimulus and increases the invariance of the representation.

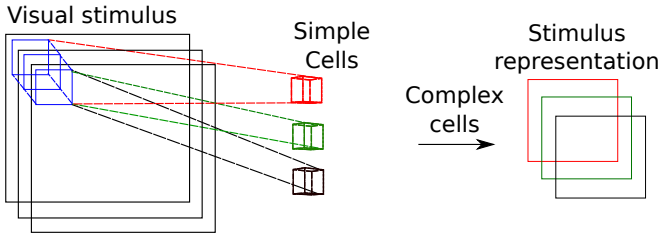

Fig. 2. Cubic receptive field implementation of the complex cells. Each filter implements a cubic kernel, which is applied to a stack of images, producing a single image which is applied to a pooling operator, simulating simple cells.

simple information and are able to generate complex feature representations of shape, orientation, position, and texture among others. Due to the different visual representations that these models could extract they were applied in several visual tasks [17], [18], [19].

In this paper we propose a CNN-based model for automatic emotion recognition. Our system is based on the visual stimuli for multimodal emotion representation described by [9] and in the deep hierarchical feature representation of the human brain described by [13]. Our system extends the visual representation of the CNN by implementing a multichannel architecture. Each channel receives different information from a sequence of visual stimuli and is able to learn and extract spatial and temporal features. The first layers extract temporal features of a sequence and pass it to deeper layers, which are able to encode complex spatial representations. Our model thus is able to learn hierarchical features, which proved to be ideal for spontaneous emotion recognition.

We evaluate our system with the CAM3D corpus of spontaneous complex mental states [20]. This corpus contains 11 different emotional states expressed spontaneously by different subjects. The emotional states present in the corpus can be applied to a range of HRI scenarios, and the evaluation of our system in this corpus extends the area of affective computing for HRI.

The paper is structured as follows: The next section introduces our multichannel convolutional neural network architecture and describes how temporal and spatial features are learned and extracted. The methodology for our experiments is given in Section II and the results are reported in Section III. A discussion about the results and the proposed system are described in Section IV. Finally, the conclusion and future work are given in Section V.

A. Learning Hierarchical Features

Our architecture implements a Multichannel Convolutional Neural Network (MCCNN) [21] to extract hierarchical features from visual stimuli. Different from a CNN, the MCCNN implements different channels, each one containing one CNN. The outputs of the CNNs are connected to a hidden layer, which is connected to a classifier. Each channel produces different and unique feature extractors after training. The first layers of each channel extract edge-like features and the deeper ones generate complex representation of the emotional expression sequence.

A CNN simulates the simple and complex cells [22], [14] by applying two operations in a CNN: convolution and pooling. The simple cells, represented by the convolution operations, use local filters to compute high-order features applying a convolution operation on the image. The complex cells extract spatial invariance by pooling simple cell units of the same time steps from previous layers for a limited range.

In a CNN, each simple cell layer has a series of different filters which are applied to the same region of the image. Each filter generates one output, resulting in several different representations of the same image for each layer. The complex cells process each of these images, generating independent rotation and scale invariance. All the representations are passed to the next layer which computes the new feature representation.

Each set of filters in the simple cell layers operates on a receptive field in the image. The activation of each unit v_{nc}^{xy} at (x,y) of the n^{th} filter in the c^{th} layer is given by

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

where $max(\cdot,0)$ implements the rectified linear function, which was shown to be more suitable than other linear functions for training deep neural architectures [23], b_{nc} is the bias for the nth feature map of the cth layer, 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 the unit (h,w) within a receptive field $c-1$, which is connected to the previous layer, 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 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 by [24]. Figure 1 illustrates the simple and complex cell processes.

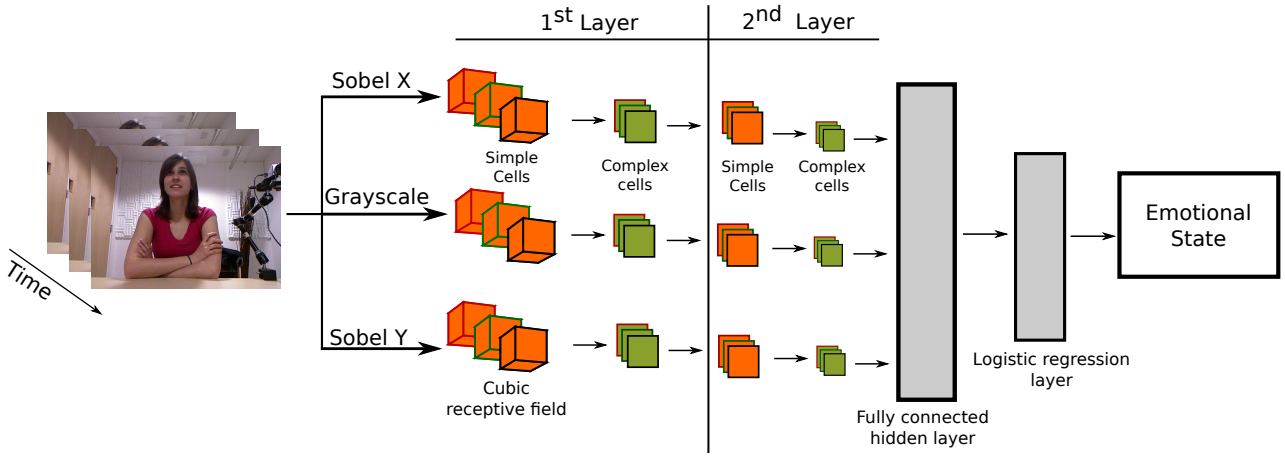

Fig. 3. Proposed architecture for a Multichannel Convolutional Neural Network using 3 channels and 2 layers. Two of the channels apply Sobel-like filters in each image of the sequence. In the first layer of each channel a cubic receptive field is implemented.

A problem shared among deep neural network architectures is the large amount of computational cost training. Usually, several layers of simple and complex cells are necessary to generate different feature representations, which increases the number of parameters to be updated during training. The multichannel implementation uses fixed filters to increase the details of features extracted on the first layers, and reduces the number of parameters during training. Our architecture uses 3 channels, each of them implementing one CNN and processing different information.

The first layer of two of the channels have Sobel-based filters before the first simple/complex cell layer. The Sobel filters are performing very simple edge enhancement and are not learned by the model. In a common CNN, the first layers will become Gabor-like filters after training. In our implementation, this representation is different, and the network is able to achieve a more complex feature representation than Gabor filters. Also, during training the three channels influence each other, driving the filters' training to a different direction than when training only one channel. The three channels share the same training process, and although the weight updates in each channel are individual, the fact that they are connected in the end creates a bias for the update. As we are applying three channels with specialized information, here represented by not trainable and constant filter maps, our architecture does not need to be so deep, which reduces the number of parameters to be updated.

B. Learning Temporal Features

An important issue for the recognition of spontaneous emotion expressions is temporal dependency. Our system creates a temporal feature representation by using a cubic receptive field implementation [25]. The cubic receptive field applies complex cells to a stream of visual stimuli. A cubic filter is applied at the same region of a stack of images. This process still extracts the spatial features of the images, but correlates between the sequences. The complex cells learn to enhance the structures which are present in the sequence, generating sequence-dependent features.

In a cubic convolution, the value of each unit (x,y,z) at the n^{th} filter map in the c^{th} layer is defined as:

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

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 $(c\text{-}1)$ 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 and z indexes each image in the image stack, R is the number of images stacked together representing the new dimension of the receptive field.

The same cubic unit is connected to the same region of a stack of images. The cubic receptive field is used only in the first layer of each channel, which is connected directly with the visual stimuli. A series of cubic receptive fields is applied to a whole image to generate different representations. For each filter, a single image is obtained containing the spatial and temporal representation. Simple cells are applied, to generate feature invariance and Figure 2 illustrates this cubic receptive field process.

Our system is trained using a limited number of images. To be able to deal with sequences with different numbers of frames, we create a sequence dependency scheme. A sequence of frames is fed into the network, and a label is generated. For each series of labels, the one occurring most often is selected.

As described in the work of Bar [26], the human brain continuously recognizes visual input based on prior knowledge that is used for a focused identification of visual stimuli. The implementation of the Sobel-filters in our architecture simulates prior knowledge by using a simple edge-like enhancement to drive the learning of features by the system. The Sobel filters represent this rudimentary

information and help the system to create a complex and specific representation of the visual stimuli accelerating the learning process. Our system uses the cubic receptive field implementation to deal with temporal features in all three channels. Due to the three channels, the system does not need to be deep and two layers are enough for the proposed scenario as we will see in Section III. Figure 3 illustrates the proposed system, with the modularization of the three channels and the cubic receptive field.

II. Methodology

To evaluate our system we used the 3D corpus of spontaneous complex mental states (CAM3D) [20]. The corpus is composed of 108 video/audio recordings from 7 different subjects in different indoor environments. Each video exhibits the upper body of one subject while an emotion is generated. 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 are labeled using crowd-sourcing: *agreeing*, *bored*, *disagreeing*, *disgusted*, *excited*, *happy*, *interested*, *neutral*, *sad*, *surprise*, *thinking* and *unsure*. Figure 4 shows examples for *agreeing*, *happy* and *thinking* sequences.

Each emotional state video present in the CAM3D corpus has varying length and sequences were recorded with different subjects. Table I shows the number of videos recorded for each emotion. The complex emotion expressions present in the CAM3D corpus can easily be applied to several HRI scenarios. It is possible to enhance a dialogue, perceiving interest, for example, or to note when a human is thinking. Perceiving spontaneous happiness or sadness can be useful when the robot needs feedback from the human in an HRI task.

Our system uses a sequence with a limited number of frames as input. To normalize the data, all the videos are separated in sub-sequences of 3 frames, generating more sequences for each emotion. We evaluate the system in a 3-fold cross validation. The same subject is not present in the videos of the training and testing subgroups at the same time. The exceptions are disgusted, excited and sad which have only one video sample. However, not all the sequences of the training subdivision are present at the testing subgroup.

The network topology used is the same in all the experiments. The network receives 3 frames as input and has 2 layers in each channel. Table II shows the network parameters

TABLE I

Number of videos available for each emotional state in the CAM3D dataset. Each video has 1 execution of the same expression.

Emotional State	Videos	Emotional State	Videos
Agreeing	4	Interested	7
Bored	3	Neutral	2
Disagreeing	2	Sad	1
Disgusted	1	Surprised	5
Excited	1	Thinking	22
Happy	26	Unsure	32

TABLE II

Parameters of the proposed system used for all experiments.

Parameters	Layer 1	Layer 2
Filters	5	10
Receptive field size	$3 \times 3 \times 3$	5×5
Sub sampling size	4×4	2×2
Neurons hidden layer	100	
Learning rate	0.01	

Fig. 4. Examples of sequences present at the CAM3D corpus. (a) shows *agreeing*, (b) being *happy* and (c) *thinking*.

used for the experiments. Each image has originally 640x480 pixels and is scaled down by a factor of 10, having 64x48 pixels before being processed by the network. Also, each image is transformed to gray scale, and the pixel intensities are normalized to have mean 0 before they are sent to the network. In this way, each receptive field in the first layer is connected to the pixel intensities of the image.

The average values of F1-Score, training, and recognition time of 30 repetitions of the experiment are computed and reported. To evaluate how important each channel of the architecture is during classification, experiments with one, two and three channels were performed. Each channel is evaluated individually and also the combination of the three channels is evaluated. All the experiments were implemented in Python using the library Theano[1] and were executed in a machine with an Intel Core 5i 2.67 Ghz processor, with 8GB of RAM.

III. Results

There are three experiments using one channel. The first one uses the raw sequence, transformed to grayscale, as input. The second one applies the Sobel filter in the X direction in each frame of the sequence and the third one the Sobel filter in the Y direction. Table III shows the average and standard deviation of the F-score, training and recognition time for experiments with one channel. Using the grayscale our system achieves a recognition rate of 85.2%, while using Sobel-X achieved 77.0% and Sobel-Y 78.1%.

[1] http://deeplearning.net/software/theano/

TABLE III

CLASSIFICATION F_1-SCORES, STANDARD DEVIATIONS, TRAINING AND RECOGNITION TIME FOR THE EXPERIMENTS USING ONE CHANNEL OF THE NETWORK.

	F-Score (%)	Training time (s)	Rec. time (s)
—Grayscale—	85.2% (+/-3.2)	67.4	0.0036
—Sx—	77.0% (+/-1.5)	66.8	0.0032
—Sy—	78.1% (+/-1.0)	66.0	0.0031

TABLE IV

CLASSIFICATION F_1-SCORES, STANDARD DEVIATIONS, TRAINING AND RECOGNITION TIME FOR THE EXPERIMENTS USING TWO CHANNELS OF THE NETWORK.

	F-Score (%)	Training time (s)	Rec. time (s)
—Grayscale+Sx—	90.8% (+/-2.0)	102.7	0.0072
—Grayscale+Sy—	91.0% (+/-1.8)	107.2	0.0068
—Sx+Sy—	82.4% (+/-4.2)	106.8	0.0064

Three experiments with two channels were evaluated. The first one contains the combination of the grayscale channel with the Sobel-X, the second one with Sobel-Y and the third one the combination of both Sobel inputs. Table IV shows the results for the two-channel experiments. The combinations of grayscale and Sobel-Y, and grayscale and Sobel-X produced a similar F-Score, around 91%. Both of the combinations with grayscale achieved a higher value than the combination of the Sobel filters, which achieved 82.4% of F1-Score.

Combining the three channels produced the highest F1-Score. Table V reports the F1-Score, training and recognition time computed for 3 channels. The F1-Score of 97.49% with 3 channels was the highest computed in all the experiments, improving by more than 6% the F1-Score reported in the two channels experiment.

IV. DISCUSSION

We are not aware of reported results using the CAM3D dataset for automatic emotion recognition, and one of the challenges of the corpus is the small number of videos for each emotion expression. However, our system was able to use the small number of sequences present in the dataset by using the limited length sequence and voting schemes. We report the first results for automatic emotion recognition for the 12 emotional states present in the corpus.

From the three experiments which our system was evaluated for, using the three channels was the one which achieved the highest F-Score, showing that the combination of the three channels produces the best feature representation. The tuning of the filters of each simple cell layer is influenced by each channel, and produced more complex features which were able to distinguish better between the emotions. When compared with the implementation of two channels, the F1-Score obtained by our system with three channels was better, but also the training and recognition time increased. Comparing our system with a common CNN implementation, used in the experiments with 1 channel, also shows that there is an increase in the F-Score of at least 22 %. When applying only the Sobel-based channels to the image, the results tend to become worse because the Sobel filter alone cannot enhance the learning of learn complex features.

TABLE V

CLASSIFICATION F_1-SCORES, STANDARD DEVIATIONS, TRAINING AND RECOGNITION TIME FOR THE EXPERIMENTS USING THREE CHANNELS OF THE NETWORK.

F-Score (%)	Training time (s)	Rec. time (s)
97.49% (+/-1.8)	186.6	0.0136

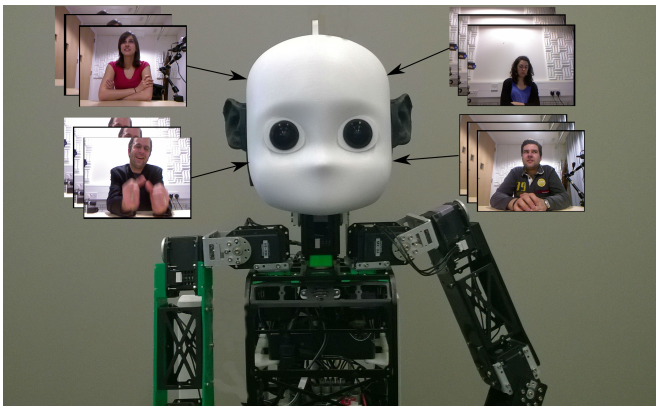

Fig. 5. Illustration of am HRI scenario where the NIMBRO robot will use complex emotional states in a learning task.

With the results, we show that our model could extract better features than a common CNN implementation, and that the use of hierarchical temporal-space features are suitable to be used for spontaneous emotion recognition.

The CAM3D corpus contains a collection of complex and spontaneously expressed mental states, which simulates the ones found in many HRI scenarios. The use of different subjects and different background increase the difficulty to use the dataset but approximates the reality of HRI indoor scenarios. Many HRI scenarios deal with robots interacting with one subject at a time, but each subject can express emotions in many different ways. Our system is able to deal with different subjects, different backgrounds and unconstrained body motion and face expressions. Also, no pre-processing step is necessary which decreases the computational effort to use it. Our system is suitable to be applied in an HRI scenario, with multiple subjects and for a spontaneously expressed complex mental state recognition. Evaluating our system with the CAM3D corpus approximates the use of our architecture in a real-world HRI scenario and gives us the robustness of a corpus established in the literature.

V. CONCLUSION

The use of emotion recognition improves how humans and robots communicate in HRI scenarios. In this paper, we propose a deep neural model for automatic spontaneous complex mental state recognition. Our system is based on

a multichannel implementation of CNN and is able to learn hierarchical features from a sequence of images, generating spatial and temporal representations of the input stimulus. Our architecture has 3 channels, each one receiving different information extracted from the same video sequence. The channels are connected to a hidden layer but are independent up to this point in the architecture.

The system is evaluated using the CAM3D corpus and is able to recognize spontaneous complex mental states with multiple subjects and different backgrounds. The corpus contains non-visual emotion expressions and it captures the upper body of the subjects. Our system is able to achieve an F1-Score of 97.5% and 3 minutes are necessary to train it. Our system achieves an F1-Score which are 20% higher than that achieved by a standard CNN. The experiments show that our architecture produces more reliable features for spontaneous emotion recognition than the standard CNN implementation.

For future work, a deeper analysis of the complex cell configurations could improve the understanding of the system. Our system will be embedded in an HRI scenario using a humanoid robot. A modified NIMBRO [27] robot with a new head and new arms will be used in a learning task and will use complex emotional states as feedback for the learning process. Figure 5 illustrates this scenario. Also, the use of temporal segmentation for emotion recognition will be studied which can improve the robustness of the system for continuous recognition.

ACKNOWLEDGMENT

The authors thank Dr. Peter Robinson for providing the CAM3D Corpus.

This work was partially supported by CAPES Brazilian Federal Agency for the Support and Evaluation of Graduate Education (p.n.5951–13–5).

The authors also would like to thank Johannes Bauer for his constructive comments and insightful suggestions that improved the quality of this manuscript.

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] T. Spexard, M. Hanheide, and G. Sagerer, "Human-oriented interaction with an anthropomorphic robot," *Robotics, IEEE Transactions on*, vol. 23, no. 5, pp. 852–862, Oct 2007.

[4] M. Cabanac, "What is emotion?" *Behavioural Processes*, vol. 60, no. 2, pp. 69 – 83, 2002.

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

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

[7] P. Rozin and A. B. Cohen, "High frequency of facial expressions corresponding to confusion, concentration, and worry, in an analysis of naturally occurring facial expressions of Americans." *Emotion*, vol. 3(1), pp. 68–75, 2003.

[8] R. Cowie, "Building the databases needed to understand rich, spontaneous human behaviour," in *Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, Sept 2008, pp. 1–6.

[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. 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, no. 810, pp. 1–9, 2013.

[11] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 39, no. 1, pp. 64–84, Feb 2009.

[12] 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.

[13] R. Adolphs, "Neural systems for recognizing emotion," *Current Opinion in Neurobiology*, vol. 12, no. 2, pp. 169 – 177, 2002.

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

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[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] S. Lawrence, C. Giles, A. C. Tsoi, and A. Back, "Face recognition: a convolutional neural-network approach," *Neural Networks, IEEE Transactions on*, vol. 8, no. 1, pp. 98–113, Jan 1997.

[18] T. P. Karnowski, I. Arel, and D. Rose, "Deep spatiotemporal feature learning with application to image classification," in *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, Dec 2010, pp. 883–888.

[19] M. Khalil-Hani and L. S. Sung, "A convolutional neural network approach for face verification," in *High Performance Computing Simulation (HPCS), 2014 International Conference on*, July 2014, pp. 707–714.

[20] 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.

[21] P. Barros, S. Magg, C. Weber, and S. Wermter, "A multichannel convolutional neural network for hand posture recognition," in *Artificial Neural Networks and Machine Learning ICANN 2014*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2014, vol. 8681, pp. 403–410.

[22] 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.

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

[24] 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.

[25] 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.

[26] M. Bar, "The proactive brain: using analogies and associations to generate predictions," *Trends in Cognitive Sciences*, vol. 11, no. 7, pp. 280 – 289, 2007.

[27] M. Missura, P. Allgeuer, M. Schreiber, C. Münstermann, M. Schwarz, S. Schueller, and S. Behnke, "Nimbro teensize 2014 team description," University of Bonn, Tech. Rep., 2014.