Emotion Recognition from Body Expressions with a Neural Network Architecture

Nourhan Elfaramawy, Pablo Barros, German I. Parisi, and Stefan Wermter
Knowledge Technology, Department of Informatics, University of Hamburg
Hamburg, Germany
{barros,parisi,wermter@informatik.uni-hamburg.de}

ABSTRACT

The recognition of emotions plays an important role in our daily life and is essential for social communication. Although multiple studies have shown that body expressions can strongly convey emotional states, emotion recognition from body motion patterns has received less attention than the use of facial expressions. In this paper, we propose a self-organizing neural architecture that can effectively recognize affective states from full-body motion patterns. To evaluate our system, we designed and collected a data corpus named the Body Expressions of Emotion (BEE) dataset using a depth sensor in a human-robot interaction scenario. For our recordings, nineteen participants were asked to perform six different emotions: *anger, fear, happiness, neutral, sadness*, and *surprise*. In order to compare our system with human-like performance, we conducted an additional experiment by asking fifteen annotators to label depth map video sequences as one of the six emotion classes. The labeling results from human annotators were compared to the results predicted by our system. Experimental results showed that the recognition accuracy of the system was competitive with human performance when exposed to body motion patterns from the same dataset.

ACM Classification Keywords

I.2.10. Vision and Scene Understanding.

Author Keywords

Emotion recognition, neural networks, learning systems.

INTRODUCTION

The ability to understand people's emotions plays a central role in human social interaction and behavior [23][24]. In the human-robot interaction (HRI) domain, research towards perceiving human emotions has recently received considerable attention supported by the fact that emotional intelligence enhances the quality of communication, interaction and social relations between people [12]. Therefore, the task of emotion recognition plays a key role in improving the interaction between humans and robots in domestic environments [9]. Artificial intelligent systems recognizing emotions and making use of affective information may significantly improve the overall HRI experience, e.g., by triggering pro-active robot behavior as a response to the user's emotional state.

Emotion recognition from body motion has received less attention in the scientific community compared to recognizing emotions from facial expressions and speech analysis. The main reason is that emotions are difficult to recognize from complex body-motion patterns. Another challenge is that people's reaction to the same emotional situation may differ since people can express different emotions based on their cultural background or habits [14], which makes it more difficult for emotions to be recognized from bodily expressions. For example, some people can react impetuously or run amok when they see a mouse, while some may stand still and cover their face with their hands. While facial expressions may more easily convey emotional information, it is often the case in HRI scenarios that a person is not facing the sensor or standing far away from the camera, i.e. facial features would be difficult to compute due to strongly degraded spatial resolution. Therefore, there is a motivation to develop an artificial system that can efficiently recognize emotions from body motion patterns.

In this paper, we propose a neural network architecture to recognize a set of emotional states from body motion patterns. The learning architecture consists of a hierarchy of self-organizing networks that learn the spatio-temporal structure of the input. In order to evaluate our system, we collected a corpus named the Body Expressions of Emotion (BEE), with nineteen participants performing six different emotional states: *anger, fear, happiness, neutral, sadness*, and *surprise*. The dataset was collected in an HRI scenario consisting of a humanoid robot Nao extended with a depth sensor to extract 3D body skeleton information in real time. In order to compare the performance of our system to human observers, we performed an additional study in which raters that did not take part in the data collection phase had to label depth map sequences as one of the six possible emotions. Our experimental results show that our system successfully learns to classify a set of training emotions from the dataset and that its performance is very competitive with respect to human observers.

HAI '17, October 17–20, 2017, Bielefeld, Germany
© 2017 ACM. ISBN 978-1-4503-5113-3/17/10. . . $15.00
DOI: https://doi.org/10.1145/3125739.3125772

In the following sections, we introduce the related work for emotion recognition using different cues such as facial features, speech, and body motion patterns. Then, we describe our neural network approach along with the details of the data collection and the evaluation of our system with respect to human observers. We conclude with a set of possible research directions to extend our current approach to a more natural HRI experience, e.g. by using multi-modal information.

RELATED WORK

The way one perceives people's emotions varies from one person to another, depending on many aspects such as personal background, belief, culture and tradition [7]. However, there are well-known human expressions that we agree on. Charles Darwin was the first one back in 1872 to categorize different facial expressions and body postures into emotions [6] which are known now as the basic affective states: *anger, fear, happiness, sadness*, and *surprise*.

In visual perception, it is argued that face expressions and body motion patterns complement each other [10]. The way our bodies move is thought be a reflection of our inner feelings and emotional states [3]. For instance, when people are angry, they might walk fast or perform aggressive gestures with their bodies. Recently, advances in computer vision and machine learning have effectively recognized faces, voices, objects, images, as well as emotions from facial and vocal expressions. However, the recognition of emotions from body expressions has received less attention. A possible issue is the conflict of views on how body motion patterns convey affective information.

Body expressions in terms of posture and motion patterns have played an important role in understanding emotions. A study by Coulson [5] showed the influence of body posture on conveying emotions through systematic analysis of static images. The experiments were conducted with observers having to classify images of body postures into emotions. Among others, the observers associated head inclination with sadness and elbow flexion with anger. Thus, the movements of head, chest, elbow and shoulder represent significant features of bodily emotions. Body motion patterns may be thought of as gestures or full-body movements that require a change of posture or body weight distribution. It has been shown that movement kinematics (e.g. velocity and acceleration) represent significant features when it comes to recognizing emotions from body patterns [25][26]. Similarly, temporal features in terms of body motion resulted in higher recognition rates than postural features alone [21]. Therefore, in our proposed approach we use both pose and motion features.

The goal of affective computing is to enable systems to recognize and express emotions in a human-like manner [23]. Emotion recognition has contributed to achieve more natural interaction between humans and robots. For instance, a socially-assistive robot may be able to strengthen its relationship with an elderly person if it can understand whether that person is bored, angry or upset. In this context, a vast amount of research has been conducted in affective states recognition in the context of HRI scenarios, e.g., using facial expressions [1][13][2], speech detection [17] or a combination of these cues [1][2].

Body expressions are thought to be more about the quantity of the emotion but not the quality since they provide information about the intensity of a specific emotion [8]. However, the problem of recognizing emotions from body motion has received significantly fewer attention [11][27][22]. A reason for this could be the fact that body emotion recognition tasks involve a higher degree of complexity and subtleties characterizing affective body expressions may be hard to extrapolate with an artificial processing system.

Kleinsmith *et al.* [11] showed significant results in addressing emotion recognition tasks from body postures. Their system based on categorizing and learning module (CALM) [16] was tested on 212 postures across 9 emotional states (angry, confused, fear, happy, interest, relaxed, sad, startled, and surprised). The experimental results have shown 70% of overall accuracy of affective postures. Schindler *et al.* [27] presented an image-based classification system for recognizing emotion from images of body postures. The overall recognition accuracy of his system resulted in 80% for six basic emotions. Although these systems show a high recognition rate, they are limited to postural emotions, which are not sufficient for a real-time interactive situation between humans and robots in a domestic environment. Piana *et al.* [22] developed a real-time emotional recognition system using postural, kinematic, and geometrical features which were extracted from sequences of 3D skeletons videos. They considered features from gestures of the following body joints: head, shoulders, elbows, hands, and torso. Their system reached an overall recognition rate of 61.3%. On the one hand, Kleinsmith *et al.* [11] and Schindler *et al.* [27] showed promising results but they were only focused on postural features of the body. On the other hand, Piana *et al.* [22] took body motion patterns into account but they did not consider full-body movements.

OUR APPROACH

We propose a neural network architecture for learning emotions from body motion patterns captured with a depth sensor. The overall view of the architecture is shown in Fig. 1, consists of a hierarchy of self-organizing networks for learning sequences of 3D body joint features. In the first layer, two Grow When Required (GWR) networks [15], G^P and G^M, learn a dictionary of prototype samples of pose and motion features respectively. In the second layer, a recurrent variant of the GWR [20], G^I, is used to learn prototype sequences and associate symbolic labels to unsupervised visual representations of emotions for the purpose of classification.

It is to be pointed out that the focus of our study is on investigating whether body expressions from depth map videos sequences convey adequate affective information for the task of emotion recognition, rather than comparing different neural network architectures for processing spatio-temporal data. (For an exhaustive comparison of recurrent GWR networks with state-of-the-art classification methods, we refer the reader to Parisi and Wermter [20].) As an advantage with respect to other well-studied methods for processing spatio-temporal data, this approach allows us to develop a learning architecture that can be trained incrementally (no re-training is needed for learning new samples or classes).

Hierarchical Learning

The use of a hierarchy of GWR networks for learning actions from body motion patterns was proposed by Parisi *et al.* [19]. The separate processing of pose and motion features and their subsequent integration has been shown to improve the topological formation of visual representations in a hierarchical learning scheme. However, this model learns sequences by concatenating neural activations from previous layers, which leads to an increasing dimensionality of the neural weights along the hierarchy. This may be an issue for high-dimensional input, e.g. the processing of sequences of body joints from 3D skeleton information. Therefore, we extended this model with a recurrent variant of the GWR, the Gamma-GWR [20], that equips each neuron in the network with a temporal context.

The learning is carried out as follows. First, the sequence of input vectors is pre-processed in order to separate the pose and motion features. Motion features are obtained by computing the difference between two consecutive frames containing pose features. Then, the two resulting datasets are sent sequentially to the GWR networks, namely G^P and G^M in the first layer. These networks are time-independent, i.e. prototype neurons learn exclusively spatial properties of the input by processing one frame at the time. After this training phase, neural activation trajectories from these two networks are concatenated as pose-motion neurons and fed into the third network, G^I, which learns the latent spatio-temporal structure of its input. This last network is also equipped with an associative mechanism to attach sample labels to the neurons.

The GWR [15] is a growing self-organizing network that learns the prototype neural weights from a multi-dimensional input distribution. It consists of a set of neurons with their associated weight vectors, and edges that create links between neurons. For an input vector $\mathbf{x}(t)$, its best-matching neuron $\mathbf{w}_b$ is computed as the neural weight that minimizes the distance to the input, such that:

$$b = \arg\min_{j \in A} \|\mathbf{x}(t) - \mathbf{w}_j\|, \tag{1}$$

where A is the set of neurons in the network. Each neuron is equipped with a firing counter η that considers the number of times that the neuron has fired. A new neuron is added if the activity of the network computed as $a = \exp - \|\mathbf{x}(t) - \mathbf{w}_b\|$ is smaller than a given activation threshold a_T and if the firing counter of $\mathbf{w}_b$ is smaller than a firing threshold f_T. This mechanism yields the creation of neurons only after the existing ones have been sufficiently trained. The neural weights and the connectivity patterns are developed following competitive Hebbian learning, thus the formation of the maps preserves the topological properties of the input. At each iteration, the neural weights are updated according to:

$$\Delta \mathbf{w}_i = \varepsilon_i \cdot \eta_i \cdot (\mathbf{x}(t) - \mathbf{w}_i), \tag{2}$$

where ε_i is a constant learning rate and the index i indicates the best-matching neuron b and its topological neighbors. The GWR has several advantages with respect to other well-known self-organization models as well as other models of growing self-organization. First, the network has the ability to add new neurons to the network whenever the current input is not sufficiently matched by the existing neurons. Second, the network learns incrementally with a fixed learning rate, which is suitable for learning dynamic input distributions. However, the standard GWR does not account for the learning of temporal input relationships.

The Gamma-GWR [20] extends the GWR with temporal context. In order to take the latent spatio-temporal structure of the input, the best-matching neuron is computed as a linear combination of the current input and K temporal context descriptors:

$$b = \arg\min_{i} \{d_i\}, \tag{3}$$

$$d_i = \alpha_w \cdot \|\mathbf{x}(t) - \mathbf{w}_i\|^2 + \sum_{k=1}^{K} \alpha_k \cdot \|\mathbf{C}_k(t) - \mathbf{c}_i^k\|^2 \ , \tag{4}$$

$$\mathbf{C}_k(t) = \beta \cdot \mathbf{w}_{b(t-1)} + (1-\beta) \cdot \mathbf{c}_{b(t-1)}^{k-1}, \tag{5}$$

for each $k = 1, ..., K$, where $\alpha_i, \beta \in (0;1)$ are constant values that modulate the influence of the current input and the past activations and $b(t-1)$ is the index of the best-matching neuron at $t-1$. The training is carried out by adapting the weight and the context vectors of the best-matching neurons and its neighbors towards the current input according to:

$$\Delta \mathbf{w}_i = \varepsilon_i \cdot \eta_i \cdot (\mathbf{x}(t) - \mathbf{w}_i), \tag{6}$$

$$\Delta \mathbf{c}_i^k = \varepsilon_i \cdot \eta_i \cdot (\mathbf{C}_k(t) - \mathbf{c}_i^k), \tag{7}$$

where ε_i is the learning rate that modulates neural update. The network activation in the Gamma-GWR is given by $a_t = \exp(-d_b)$ with d_b as defined in Eq. 4. In this network, we set $K = 9$ so that it takes into consideration the last 10 input frames, i.e. the current input plus 9 previous neural activations. The training parameters are discussed in the following section.

Classification

The aim of the classification process is to predict the emotion class of unseen data samples after the training phase. In our case, the labels are represented by the six emotion classes. The learning process of the Gamma-GWR is unsupervised, i.e. sample labels are not used for the formation of visual representations. Therefore, we associate symbolic labels with neural weights during the learning process in G^I as proposed in [20]. After the training phase, it is expected that each neuron in G^I is attached to the label that better represents the activation of that neuron according to labels of the training set.

During the prediction phase, unlabeled novel samples are processed by the hierarchical architecture, yielding patterns of neural weight activations. For 10 processed input frames, one best-matching neuron in G^I will activate according to Eq. 3, so that the predicted label consists of the label attached to that neuron.

EXPERIMENTAL RESULTS

We now describe the Body Expressions of Emotion (BEE) dataset, the labeling process, and the evaluation of emotion recognition of our system with respect to human performance.

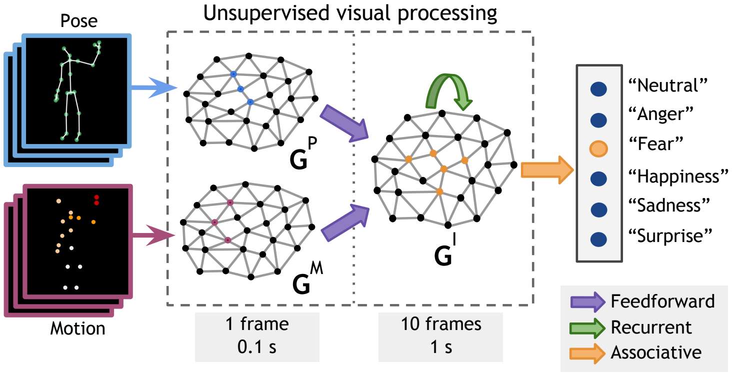


Figure 1. Our neural learning architecture with a hierarchy of self-organizing networks. The first layer processes separately pose and motion features from individual frames, whereas in the second layer a recurrent network learns the spatio-temporal structure of the joint pose-motion representations.

BEE Dataset

We collected a data corpus consisting of six affective states: *anger, fear, happiness, neutral, sadness*, and *surprise*. Nineteen participants took part in the data recordings (fourteen male, five female, age ranging from 21 to 33). The participants were students at the University of Hamburg, Germany and all of them declared not to have suffered any physical injury resulting in motor impairments. Participants came from nine different countries, resulting in multiple ways of expressing body emotions according to different cultural backgrounds.

The HRI set-up consisted of two humanoid robots Nao standing on a table. One of the robots had a depth sensor Asus Xtion mounted on top to capture depth map video sequences (Fig.2). Experiments started with participants facing this Nao at a distance of 3.5 meters. A second Nao was placed on the side and used only for the purpose of interaction so that the person will be captured from both a frontal and a side view. The sensor captured the participant while walking from the front, the back, as well as the side. First, emotionally neutral body motion was recorded to serve as a baseline. Then, the six emotions were recorded using different scenarios. In order to trigger spontaneous behavior during their performance, participants were described a brief scenario to take inspiration from. For instance, in the case of Fear participants were told: "Someone has just broken into your apartment and you are very scared. Walk slowly towards the robot so that no one will hear you and ask him to call the police." Each participant performed the same emotion five times, for a total of 30 sequences per participant (570 in total).

The use of a depth sensor has two main advantages: i) the detection of body motion is robust to light changes and ii) it allows to estimate a 3D skeleton model in real time for the subsequent processing of body motion patterns (see Fig. 3). From the depth map videos, we extracted the 3D skeleton information using OpenNI/NITE [1] and pre-processed 11 body joints of interest: head, neck, torso, left shoulder, right shoulder, left elbow, right elbow, left hand, right hand, left hip, and right hip. These joints were selected based on their influence in conveying emotions from body motion [5]. Body joints were captured using the absolute coordinates with respect to the sensor's position. In order to yield invariance to translation, we post-processed all the joints using the coordinates of the torso as the reference point. Videos were captured with a 640×480 image resolution at 30 frames per second. To reduce noise, we computed the median every 3 frames, thus resulting in body motion sequences at 10 frames per second.

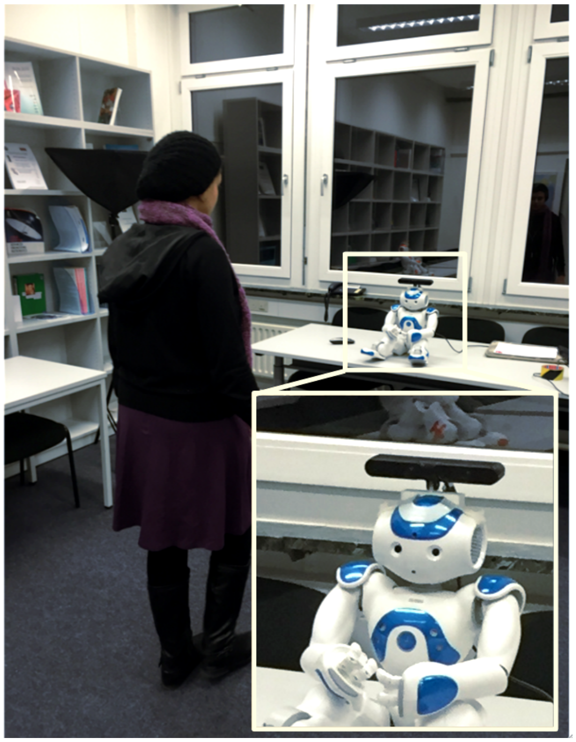


Figure 2. The data collection set-up with a humanoid robot Nao and a depth sensor.

Neural Network Training

To evaluate the classification accuracy of our system, we trained on 60% of the video sequences and tested on the remaining 40%. We empirically defined a set of neural network parameters that yielded the best classification accuracy. For the three networks G^P, G^M, and G^I we used the training pa-

[1] OpenNI/NITE: https://structure.io/openni

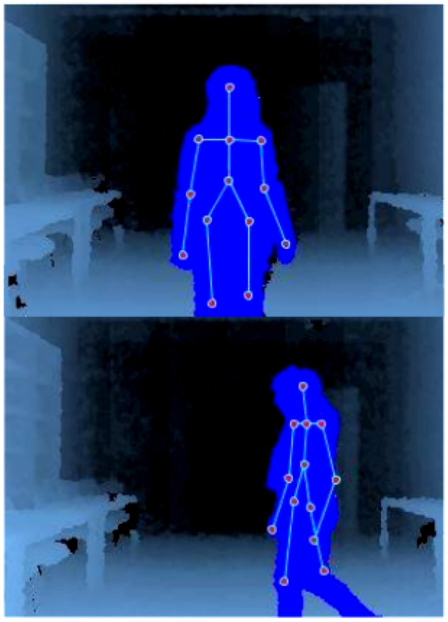


Figure 3. : A side view (right) and a frontal view (left) of the detected skeleton with eleven body joints.

Parameter	*Value*
Insertion threshold	$a_T = 0.9$
Firing threshold	$f_T = 0.1$
Learning rates	$\varepsilon_b = 0.1, \varepsilon_i = 0.001$
Training epochs	350

Table 1. Training parameters for the hierarchical neural architecture.

rameters listed in Table 1. Additional GWR parameters were set according to recommended values in the literature [15].

For the recurrent network G^I, we set $K = 9$ (i.e. each neuron is activated by a sequence of 10 frames), $\beta = 0.7$, and an activation leaky integrator $\alpha_i = [0.3, 0.2, 0.1, 0.08, 0.06, 0.05, 0.04, 0.03, 0.02, 0.01]$. It has been shown that $\beta \in [0.6, 0.7]$ and decreasing values of α_i yield minimal temporal quantization error [18]. After the training, these parameters resulted in networks having the following number of neurons $G^P = 6990$, $G^M = 3316$, and $G^I = 3878$.

Evaluation

After the training session, the overall accuracy of the proposed Gamma-GWR model on the BEE dataset is 88.8%. The average classification accuracy, precision, recall, and f-score are shown in Fig. 4. We also report a confusion matrix (Fig. 5). To be pointed out is that the correct classification of emotions is subject to a number of factors which go beyond the properties of the neural network mechanism used to predict emotion classes from unseen samples. In fact, the neural network architecture learns the latent spatial and temporal structure of the dataset, and misclassification may be due to the fact that the body motion patterns are complex and vary significantly from subject to subject.

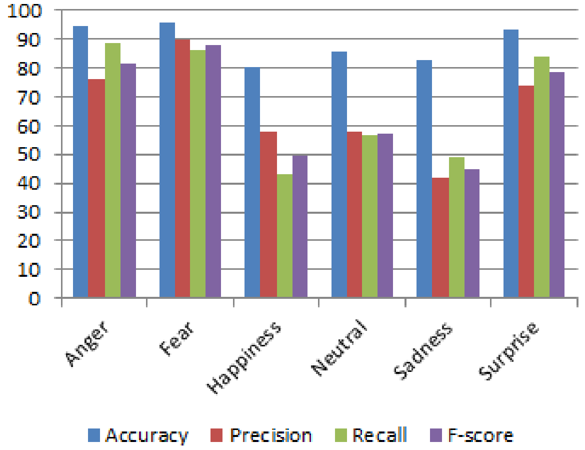
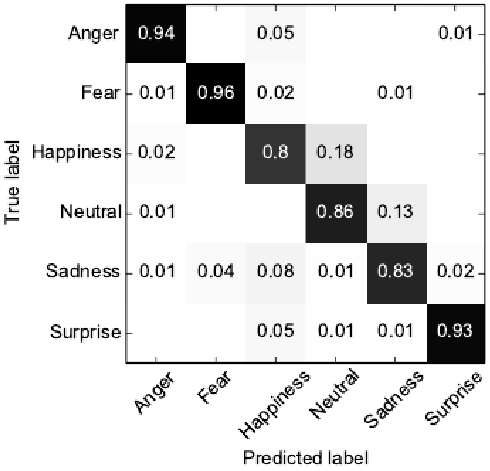


Figure 4. Average classification accuracy, precision, recall, and f-score for all the emotions.



	Anger	Fear	Happiness	Neutral	Sadness	Surprise
Anger	0.94		0.05			0.01
Fear	0.01	0.96	0.02		0.01	
Happiness	0.02		0.8	0.18		
Neutral	0.01			0.86	0.13	
Sadness	0.01	0.04	0.08	0.01	0.83	0.02
Surprise			0.05	0.01	0.01	0.93

Figure 5. The confusion matrix for the proposed Gamma-GWR model on the BEE dataset.

In order to qualitatively evaluate the accuracy of our model, we compared our system with human performance. For this purpose, we conducted an additional study in which we asked a number of people that did not take part in the data collection to label depth map video sequences as one of the six emotion classes. Fifteen raters took part in this data labeling study (thirteen male and two female, age ranging from 21 to 37). The participants were asked to watch depth map sequences on a desktop monitor and label each sequence according to the six possible emotion classes. For each participant, each emotion was displayed five times.

The final ground-truth label of emotion perception was given according to the most voted emotion class by the human observers. The majority of observers agreed on the labeling of the expressed emotion as they were intended during our data collection. We measured the intraclass correlation coefficient (ICC) [4], which indicated the inter-rater reliability of quantitative data or the correlation between different raters voting the same given subjects. In our case, the raters were the fifteen participants who labeled the emotions from videos

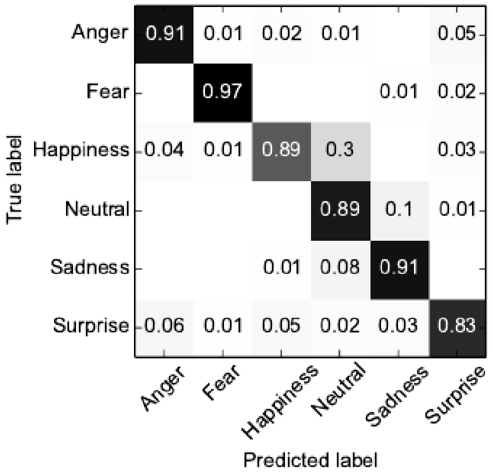


Figure 6. The confusion matrix for human observers on the BEE dataset.

and the subjects were the six different emotions. Since we had consistent raters throughout the labeling process, we applied the two-way random ICC. In this model, each subject was assessed by a different rater with the raters being randomly selected, assuming that both were randomly drawn from larger populations. The following equation was used to measure the ICC:

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}, \tag{8}$$

where σ_b^2 indicates the variance of the trait between subjects, σ_w^2 is the pooled variance within subjects, with $\sigma_b^2 + \sigma_w^2$ being the total variance of ratings (i.e. the variance for all ratings, regardless of whether they are for the same subject or not).

Our obtained average intraclass correlation coefficient was 96.5%, which indicates a high correlation between raters when exposed to the same subjects. In other words, the reliability of their observations for recognizing the six emotions is very high. A comparison between the results obtained by our system and the human participants is summarized in Table 2, showing competitive results. The overall accuracy of emotions recognized by human observers was 90.2% (whereas our system showed an overall accuracy of 88.8%).

We now analyze the recognition of each emotion from both the human and the system perspectives. In order to gain insight into the average recognition error, we computed the percentage error according to each emotion. Recognition errors show that humans could not recognize *surprise* accurately, with a recognition error of 73.3% compared to 26% of error in our system. Moreover, there was also a huge difference in recognition for the *neutral* state which was recognized more accurately by humans (93%) with respect to our system (86%). However, comparing the confusion matrix of our proposed system (Fig. 5) and the one for the human observers (Fig. 6), it can be seen that there is no relation in the way that the model and humans misclassified emotions. Additionally, this comparison is indicative and our proposed experiments are subject to a number of limitations.

	System	Human
Accuracy	88.8%	90.2%
Precision	66.3%	70.1%
Recall	68%	70.7%
F-score	66.8%	68.9%

Table 2. A comparison of overall recognition of emotions between our system and human performance.

First, the collected recordings were acted, which made the expressed emotions not spontaneous and may be exaggerated in some cases. The possible improvement could be used to conduct recordings in a real-world domestic environment. Another concern was during the data labeling phase. The videos were displayed to observers were in depth images. However, some people found difficulties in recognizing the expressed emotion in the videos from such images. This issue may be addressed by showing the videos as RGB images while covering the face of the performers.

However, the use of RGB images would also compromise the experiment since one could argue that, in this case, human observers may use facial cues to facilitate the recognition of emotions, while the purpose of our study was to assess the capability of our system to recognize emotions from body expressions only. A way to extend this experimental design is to investigate the interplay of body and facial cues in emotion recognition. We expect the accuracy of the system to increase significantly if both 3D skeletons and face expressions are used to classify emotions. In this case, the system may use either cue if one is missing or combine them to solve ambiguities.

CONCLUSION

Artificial systems recognizing emotions and making use of affective information may significantly improve the overall HRI experience by triggering pro-active robot behavior as a response to the user's emotional state. The recognition of emotional states from body motion patterns can be particularly useful in HRI scenarios. First, body expressions may convey an additional social cue to reinforce or complement facial expressions. Second, this approach may be preferred when the user is not facing the sensor at an adequate distance required to effectively process facial features. We collected a dataset of six standard emotions and proposed a self-organizing neural architecture to learn and classify emotional states from 3D skeleton information. The obtained results show that our system compares with human-like performance for the recognition of emotions from full-body cues only.

The current study may be extended in different directions. First, in our study we used skeleton information from depth map video sequences, however, a self-organizing neural approach have shown state-of-the-art results for learning body features from raw video containing body motion [20]. Second, we could investigate the development of a multi-modal emotion recognition scenario, i.e. by taking into account auditory information that complements the use of visual cues [2]. The integration of audio-visual stimuli for emotion recognition has been shown to be very challenging but also strongly promising for a more natural HRI experience.

ACKNOWLEDGMENTS

This research was supported by the German Research Foundation (DFG) under project Transregio Crossmodal Learning (TRR 169), the SECURE European Training Network, and the Social Cognitive Robotics in the European Society (SOCRATES) project.

REFERENCES

1. Fernando Alonso-Martin, Maria Malfaz, Joao Sequeira, Javier F. Gorostiza, and Miguel A. Salichs. 2013. A Multimodal Emotion Detection System during Human-Robot Interaction. *Sensors* 13, 11 (2013), 15549–15581.
2. Pablo Barros and Stefan Wermter. 2016. Developing crossmodal expression recognition based on a deep neural model. *Adaptive Behavior* 24, 5 (2016), 373–396.
3. I. Bartenieff and D. Lewis. 1980. *Body movement. Coping with the environment.* Psychology Press.
4. Domenic V. Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 6, 4 (1994), 284–290.
5. Mark Coulson. 2004. Attributing Emotion to Static Body Postures: Recognition Accuracy, Confusions, and Viewpoint Dependence. *Journal of Nonverbal Behavior* 28, 2 (2004), 117–139–139.
6. Charles Darwin. 1872. *The expression of the emotions in man and animals*. London, John Murray.
7. Paul Ekman and Wallace V. Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17, 2 (1971), 124–129.
8. P. Ekman and W. V. Friesen. 2001. Head and body cues in the judgment of emotion: A reformulation. *Perceptual and motor skills* 24, 3 (2001), 711–724.
9. Michael A. Goodrich and Alan C. Schultz. 2007. Human-robot Interaction: A Survey. *Found. Trends Hum.-Comput. Interact.* 1, 3 (2007), 203–275.
10. Y. Gu, X. Mai, and Y. J. Luo. 2013. Do bodily expressions compete with facial expressions? Time course of integration of emotional signals from the face and the body. *PLoS ONE* (2013).
11. A. Kleinsmith, T. Fushimi, and N. Bianchi-Berthouze. 2005. An incremental and interactive affective posture recognition system. In *In International Workshop on Adapting the Interaction Style to Affective Factors*. 378–387.
12. P.N. Lopes, P. Salovey, and R. Straus. 2003. Emotional intelligence, personality, and the perceived quality of social relationships. *Personality and Individual Differences* 35, 3 (2003), 641–658.
13. Muharram Mansoorizadeh and Nasrollah Moghaddam Charkari. 2010. Multimodal information fusion application to human emotion recognition from face and speech. *Multimedia Tools and Applications* 49, 2 (2010), 277–297.
14. Hazel Rose Markus and Shinobu Kitayama. 1991. Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review* (1991), 224–253.
15. Stephen Marsland, Jonathan Shapiro, and Ulrich Nehmzow. 2002. A self-organising network that grows when required. *Neural Networks* 15, 8-9 (2002), 1041–1058.
16. Jacob M.J. Murre, R. Hans Phaf, and Gezinus Wolters. 1992. CALM: Categorizing and learning module. *Neural Networks* 5, 1 (1992), 55–82.
17. Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. 2003. Speech emotion recognition using hidden Markov models. *Speech Communication* 41, 4 (2003), 603 – 623.
18. German Ignacio Parisi, Sven Magg, and Stefan Wermter. 2016. Human Motion Assessment in Real Time Using Recurrent Self-Organization. In *Proc. of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), New York, US*. 71–76.
19. German Ignacio Parisi, Cornelius Weber, and Stefan Wermter. 2015. Self-Organizing neural integration of pose-motion features for human action recognition. *Frontiers in Neurorobotics* 9, 3 (2015).
20. German Ignacio Parisi and Stefan Wermter. 2017. Lifelong Learning of Action Representations with Deep Neural Self-Organization. In *The AAAI 2017 Spring Symposium on Science of Intelligence: Computational Principles of Natural and Artificial Intelligence, Standford, US*. 608–612.
21. A. Patwardhan and G. Knapp. 2016. Multimodal Affect Recognition using Kinect. *arXiv:1607.02652* (2016).
22. S. Piana, A. Stagliano, F. Odone, A. Verri, and A. Camurri. 2014. Real-time automatic emotion recognition from body gestures. *arXiv:1402.5047* (2014).
23. Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA, USA.
24. Robert Plutchik. 1991. *The Emotions*. University Press of America.
25. Frank E. Pollick, Helena M. Paterson, Armin Bruderlin, and Anthony J. Sanford. 2001. Perceiving affect from arm movement. *Cognition* 82, 2 (2001), B51–B61.
26. Misako Sawada, Kazuhiro Suda, and Motonobu Ishii. 2003. Expression of Emotions in Dance: Relation between Arm Movement Characteristics and Emotion. *Perceptual and Motor Skills* 97, 3 (2003), 697–708.
27. Konrad Schindler, Luc Van Gool, and Beatrice de Gelder. 2008. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural Networks* 21, 9 (2008), 1238 – 1246.