Learning Human Behavior for Emotional Body Expression in Socially Assistive Robotics

Nguyen Tan Viet Tuyen, Sungmoon Jeong, Nak Young Chong
School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan
E-mail: {ngtvtuyen, jeongsm, nakyoung}@jaist.ac.jp

***Abstract*—Generating emotional body expressions for socially assistive robots has been gaining increased attention to enhance the engagement and empathy in human-robot interaction. In this paper, we propose a new model of emotional body expression for the robot inspired by social and emotional development of infant from their parents. An infant is often influenced by social referencing, meaning that they perceive their parents' interpretation about emotional situations to form their own interpretation. Similar to the infant development case, robots can be designed to generate representative emotional behaviors using self-organized neural networks trained with various emotional behavior samples from human partners. We demonstrate the validity of our emotional behavior expression through a public human action dataset, which will facilitate the acquisition of emotional body expression of socially assistive robots.**

***Keywords*—human-robot interaction, emotional body expression, imitation learning, clustering.**

1. Introduction

Human body expressions play an important role in non-verbal communication to facilitate the recognition of emotions. Psychological researches have shown that human emotions can be reflected in body languages and facial expressions during social interactions [1]. In recent years, many researches focused on generating emotional body expressions by estimating and incorporating the emotional state of robots, which is believed to increase the engagement and empathy between humans and robots [2]. MIT's Kismet [3] can show different facial expressions depending on valance, arousal and stance values which represent its emotional state. The facial expression of Kismet was inspired by the idea of psychological research on human facial expression [4]. From the persepectives in social psychology, body languages play an important role not only for the recognition of emotions but for emotional expression [5]. For that reason, a lot of attention has been paid to generate emotional body expressions for socially assistive robots. Emotional expression with body movement and eye color for NAO robot was proposed by Markus [6]. That paper was also motived by psychological researches about the connection between emotion and body movement, sound and eye color [7]. Based on psychological approaches to emotional body expression [1][8], classical emotional body expressions could be simply generated for SoftBank's humanoid robot Pepper as shown in Fig. 1.

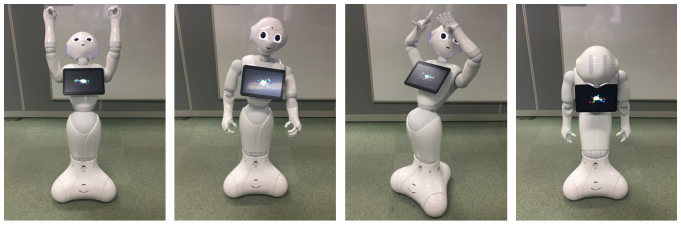

Fig. 1. Emotional Expression of Pepper robot: Happy, Relaxed, Afraid, Sad

On the other hand, in order to increase the engagement of the conversation and the empathy between a robot and a human during daily interactions, careful attention should be paid to generate an appropriate emotional robot expression according to the personality and cultural identity of a person. Therefore, robots should be required to generate their emotional behaviors in alignment with the emotional state of a person. To achieve this goal, this paper is inspired by infant's social development from a psychological point of view. According to psychological researches, one of the most common things that humans do is gathering their interested information from the surrounding environment and then utilizing it in order to form their own interpretation and actions [9]. Human behavior is often influenced by social referencing, meaning that humans tend to use the perception of another person's interpretation in order to form their own interpretation about specific situations [10]. That is the way how infants acquire the new interpretations for their social development. An interesting example was mentioned in [10], where a 9 month old infant sees that his father plays with a novel toy. The infant infers that his father likes the toy because he smiles. Then, the infant may assimilate this favorable interpretation which can influence her/his own behavior when given an opportunity to play with the toy in the future. The infant's social development was an interesting motivation for this paper to generate emotional body expressions of socially assistive robots, allowing the robots to enter into natural and intuitive social human-robot interaction. In order to achieve this goal, this paper propose that robots should pay attention to their owner's emotional body expression associated with a specific emotional state. To this end, robot generates an representative emotional expression by considering two steps as: (1) clustering of human emotional behavior samples into different groups based on similarity of body movement and (2) utilizing the most frequently observed behavior as the reference for generating robot's emotional body expression.

In this paper, we review the related literatures based on psychological researches about the connection between emotion and body or facial expressions. Then, we introduce our new approach for generating robot emotional body expressions which was inspired by infant social development. In the Methodology part, we describe how the robot acquires knowledge about human emotional body expressions as the reference information. In the Experiments and Results part, a public data set was used to evaluate our approach. A discussion about experimental results and future works was mentioned in the Discussion part and also summarize the results as well as our future work in the Conclusion and Future Work part.

2. METHODOLOGY

In order to obtain human body expression information, Kinect sensor can be easily used to extract demonstrator's skeleton. A set of actions $A_1, A_2, ..., A_n$ were gradually received during day by day human-robot interaction. Action $A_i = [S_1, S_2, ..., S_T]$ is the sequence of frames over a period of time T and $S_t = [x_1, x_2, ..., x_{20}; y_1, y_2, ..., y_{20}; z_1, z_2, ..., z_{20}]$ is the human skeleton information including 20 joint positions at time t. To extract a feature vector, the Covariance Descriptor method was used to encode sequence of frames A_i. Because an representative emotional behavior should be defined among a set of action samples, a generative model is an appropriate approach for clustering a set of human emotional body expression $A_1, A_2, ..., A_n$ into j clusters by considering the distribution of body movements. Then robot can utilize each cluster as a reference information for generating its emotional body expression by observing a human partner. This approach will be represented in part 2.2.

2.1. Covariance Descriptor

An appropriate method should be applied to create feature vectors for sequence of skeleton motion during running time. The simple approach using skeletal joint angles, joint angle velocities and velocity of joints was proposed by Fothergill [11]. However, number of frames should be equal to create the same length of feature vectors. On the other hand, Hussein et al [12] proposed a novel covariance descriptor approach which encoded the temporal dependency of joint locations. The dimension of Covariance descriptor is fixed and it is independent from the number of frame sequences. Moreover, the accuracy outperformed the state of art in human action recognition [13].

Consider an human action which is performed over T frames and $S = [x_1, x_2, ..., x_k; y_1, y_2, ..., y_k; z_1, z_2, ..., z_k]$ is the vector of skeleton joint positions at a frame t, vector S represented for K joints of skeleton, as the result, $N = 3 \times K$ elements were included in vector S. The covariance matrix is computed by

$$C(S) = \frac{1}{T-1} \sum_{t=1}^{T} (S - \overline{S})(S - \overline{S})' \quad (1)$$

where $\overline{S}$ is the sample mean of S and the $'$ is the transpose operator. The covariance descriptor was extracted from upper triangle of $C(S)$ including $N \times (N+1)/2$ elements.

2.2. Self Organizing Map

From a set of descriptors which represents for set of human actions, in this step, clustering of human actions into different groups should be carried out to find a representative human actions. For that reason, this paper proposed to use Self Organizing Map (SOM) for clustering. SOM was originally introduced by Kohonen in 1998 [14] which creating a set of neurons representing for the distributions of whole original dataset. SOM ensures the topological properties of the descriptors were preserved after reducing from d-dimensional input space to low-dimensional space grid. Meaning that, if two different behavior samples were closed to each other in an original feature space, they should be remained with similar topological property in different dimensional grid. 2-dimensional grid are usually used as a suitable visualization surface for showing similarity between features. However, it should be emphasized that this visualization can only utilize the qualitative information about original dataset [15]. In order to automatically cluster a set of descriptors, this paper firstly used SOM method for designing a grid of neurons which was considered as the training phase. Then, at the second phase, those neurons were clustered into different groups by using Distance Matrix Based approach [16]. Those descriptors in the original data and its corresponding neurons were finally found out based on Best Matching Unit (BMU) technique.

2.2.1 SOM Training Phase

For the n input descriptors, each descriptor $x_i = [x_{i1}, x_{i2}, ..., x_{id}]$ included d-dimensional features, a grid of $p \times r$ neurons was defined, each neuron represented by a prototype vector $m_i = [m_{i1}, m_{i2}, ..., m_{id}]$. During the training time, an input sample x_{sample} was picked up, then the wining neurons m_{wining} with the shortest distance to the x_{sample} was defined by BMU:

$$||x_{sample} - m_{wining}|| = min\{||x_{sample} - m||\} \quad (2)$$

The wining neuron m_{wining} was updated to make them move closer to the input sample x_{sample} with the highest intense comparing with the rest of neurons by the equation below:

$$m_{wining} = m_{wining} + \alpha(t) \times (x_{sample} - m_{wining}) \quad (3)$$

Where $\alpha(t)$ is the learning rate at time t.

It should be noticed that not only the wining neurons are updated with the new weight value, but neighbors of wining node m_i are also affected through the neighborhood kernel function $\phi(m_i, m_{wining})$ by

$$m_i = m_i + \alpha(t) \times \phi(m_i, m_{wining}) \times (x_{sample} - m_i) \quad (4)$$

Neighborhood kernel function indicates the intensity of wining neuron affecting on its neighborhood. In this paper, we used Gaussian kernel function since the global topological

relationship was better preserved [17]. Meaning that, the grid of neurons effectively reflects the distribution property of the original data. The advantage of topological preservation played crucial role for the second phase-clustering the neurons.
The Gaussian kernel function was written as:

$$h_{wi}(t) = exp\left(-\frac{||r_{wining} - r_i||^2}{2\sigma^2(t)}\right) \quad (5)$$

where r_{wining} and r_i are the location of the wining neuron and the neuron i on the grid map, respectively.

2.2.2 SOM Clustering Phase

At the second phase - clustering neurons into separated clusters, several approaches were suggested such as agglomerative clustering or k-means algorithm [15] [18]. By using k-means for clustering neurons, this involved making several k-means clustering trials with different values of k [15] and the best clustering should minimize the value of Davies-Bouldin index [19]. However, the minimum value of Davies-Bouldin index was not always indicating the appropriate number of clusters. Another method is using distance matrix for clustering [16]. Distance matrix utilizes the advantage of SOM which the topological properties of the original data were preserved after training phase, as the result, distance between neighboring neurons are approximately proportional to the distribution of the original data [16]. In this method, local minima of the grid of neurons, called representative local neurons, can be selected by

$$f(m_i, N_i) \leq f(m_j, N_j) \quad \forall j \in N_i \quad (6)$$

where $f(m_i, N_i) = median\{||m_i - m_j||\}$ is the median distance between neuron i and its neighboring neurons j. After the representative local neurons are defined by Eq. (6), all neurons are appropriately clustered by minimizing the distance between the representative local neurons and them. Since neurons are designed to encode original data into a set of Voronoi sets $V_i = \{x|||x - m_i|| \leq ||x - m_j||\forall j \neq i\}$, as the result, each neuron and its corresponding data was defined by BMU function:

$$||x - m_i|| = min\{||x - m||\} \quad (7)$$

3. Experiments and Results

3.1. Dataset

To validate the proposed model, the public Microsoft Research Cambridge-12 Kinect gesture dataset (MSRC-12) [11] was used. The dataset which includes 12 different gestures and 20 joint positions information of each gestures. In this experiment, 1640 gesture instances from 15 subjects are used. We assume that these actions were acquired from human emotional expression during daily human-robot interaction in the same emotional situation as happy. Therefore, robot can use these dataset to generate a representative emotional behavior as happy by classifying individual person's actions into different clusters. The clustering result was utilized as references for robot generating happy expression.

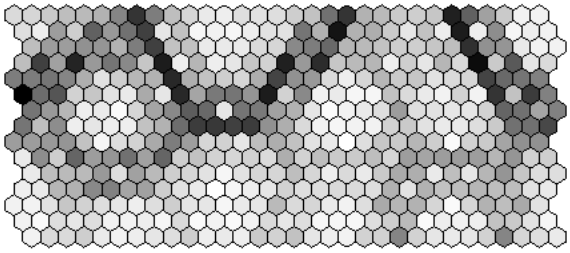

Fig. 2. U-matrix from original action dataset of subject 13

3.2. Evaluation Criteria

In this paper, precision, Recall and F-value were used to evaluate our experimental performance. These values are defined as below:

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$F_{value} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (10)$$

where True Positive (TP) represents the performance that similar actions are located in the same cluster, False Positive (FP) represents the performance that dissimilar actions wrongly clustered as a same cluster, and False Negative (FN) is the performance that similar actions are clustered by different class.

3.3. Experimental Results

A set of covariance descriptors was used for training a gird of neurons in SOM training phase. The Euclidean distance between each neurons and its corresponding neighbor were visualized in 2 dimensional feature space by U-matrix as shown in Fig. 2.

After finished the SOM training, a grid of SOM neurons was clustered into appropriate clusters based on distance between selected local representative neurons and neighborhood neurons. A set of representative neurons was selected by Eq. (6) to find a local minimum value in a grid of SOM neurons as shown in Fig. 3. Hence, number of selected representative neurons equals to the number of clusters in grip map. Finally, all neurons are assigned into appropriate clusters as shown in Fig. 4.

As mentioned in part 2.2, topological properties of the feature descriptors were preserved on the grid of SOM neurons. On the other hand, each neuron created a Voronoi region in the original feature descriptor space. As the result, each neuron and its corresponding feature descriptors were defined by BMU Eq. (7). Feature descriptors were assigned into the same clusters if its corresponding neurons located in the same clusters.

Figure 5 presents the example of TP of clustering result by using 13th subject's data. In order to evaluate the proposed model using entire dataset, we calculate the average values

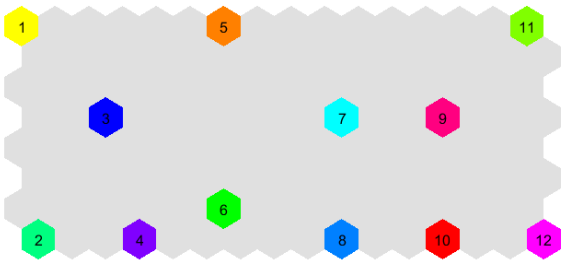

Fig. 3. 12 local minima neurons were detected from grip of SOM neurons which represent for action dataset of subject 13

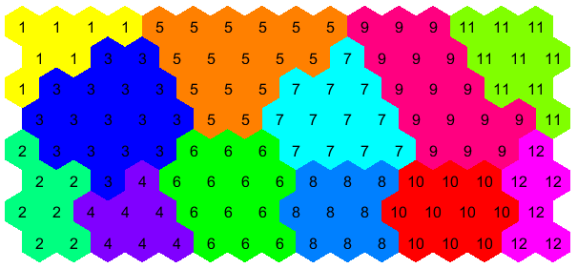

Fig. 4. 12 clusters were detected from grip of SOM neurons which represent for action dataset of subject 13

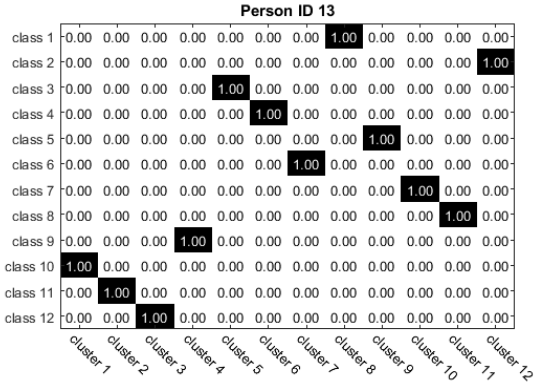

Fig. 5. Confusion matrix of True Positive (TP) by using 13th subject's data

of $Precision$, $Recall$ and F_{value} using 15 subjects data as shown in Table 1.

TABLE 1
THE AVERAGE VALUES AFTER 15 EXPERIMENTS

	percent
$Precision$	0.9166
$Recall$	0.9115
F_{value}	0.9133

4. DISCUSSION

To analyze the proposed model in more detail, we use 13th person's data to generate the confusion matrix as shown in Fig. 6 and the data was assigned into a specific $cluster$ j. It was obvious that among 12 different clusters, $cluster$ 5 is the largest one with 14 similar actions $A_1, A_2, ..., A_{14}$ located inside. The largest cluster means the most frequent body expression for happy emotion which the user always performed during human-robot interaction. As the result, robot should imitate an action $A_i \in \{A_1, A_2, ..., A_{14}\}$ which located nearest from the center of $cluster$ 5 as the representative body expression for happy emotion. Action A_i contained a set of frames $A_i = [S_1, S_2, ..., S_T]$ which $S_T = [x_1, x_2, ..., x_{20}; y_1, y_2, ..., y_{20}; z_1, z_2, ..., z_{20}]$ represent for human joint positions at time T. A transfer algorithm which converts from human joint positions S_T to Pepper robot joint angles θ_T will be investigated in our future work. It should be emphasized that the kinematic models between human and Pepper robot are different and the degree of freedoms (DOFs) as well as range of joint angles in terms of Pepper robot are limited to compare with a human model. Therefore, imitation approaches are required to satisfy physical constraints of the Pepper robot, at the same time, the meaning of human emotional expression A_i is preserved after mapping to robot model.

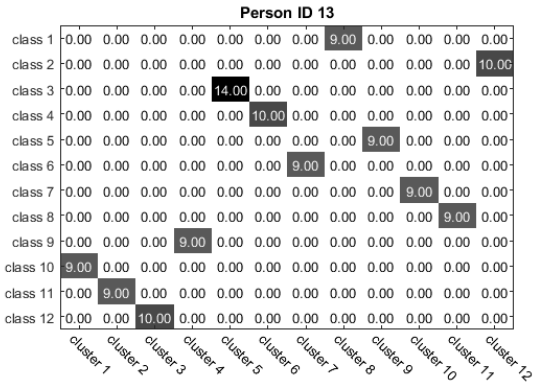

Fig. 6. Confusion matrix representing for number of actions belong to $class$ i were assigned into $cluster$ j from subject 13

Neurons and its corresponding feature descriptors were defined by BMU approach as mentioned above. However, it was noticed that, the number of clusters on the grid of SOM neurons and on the original feature descriptors were not always the same. The reason was that there were no feature descriptors located in the Voronoi regions which were created by corresponding neurons in that clusters. As shown in Fig.7, there were 13 clusters created by a grid of SOM neurons representing for action data set of subject 8. The corresponding confusion matrix in Fig.8 indicated that there were actually just 12 clusters were created since there were no elements located in $cluster$ 12.

Because MSRC-12 dataset, which is gathered by Kinect sensor, was much more noisy than HDM05-MoCap dataset [20] from motion capture sensors [21], same action class can be divided into many sub-clusters. That problem happened in dataset of subject 3 as shown in Fig. 9 when action $class$ 2 and action $class$ 7 were divided into sub-clusters respectively. The experiment also noticed same problem occurred in subject

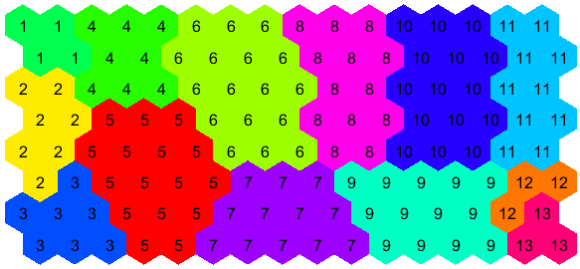

Fig. 7. 13 clusters were detected from grip of SOM neurons which represent for action dataset of subject 8

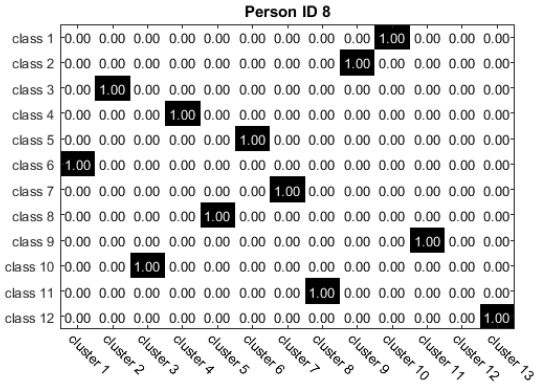

Fig. 8. Confusion matrix after clustering actions into different clusters of subject 8

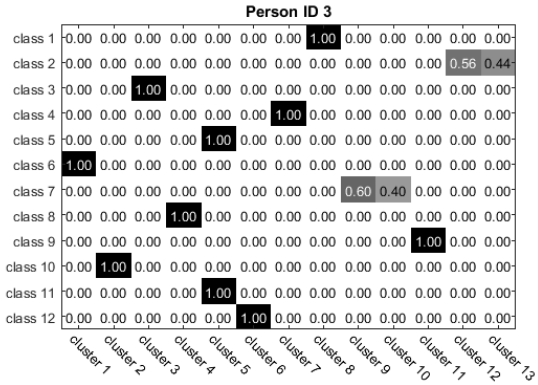

Fig. 9. Confusion matrix after clustering actions into different clusters of subject 3

7 as shown in Fig.10 which actions $class$ 6 were divided into $cluster$ 7 and $cluster$ 9 respectively.

MSRC-12 dataset includes 12 different actions classes. During the experiment, it was noted that the actions $class$ 5 and actions $class$ 11 were sometimes assigned into the same cluster. In dataset of subject 3, 100% actions $class$ 5 and 100% actions $class$ 11 were assigned into the same $cluster$ 5 as shown in Fig.9. The same problem happed at action dataset of subject 7 when both actions $class$ 5 and $class$ 11 were located in them same $cluster$ 2 as shown in Fig. 10. Actions $class$ 5 were described as the movement of both arms in front

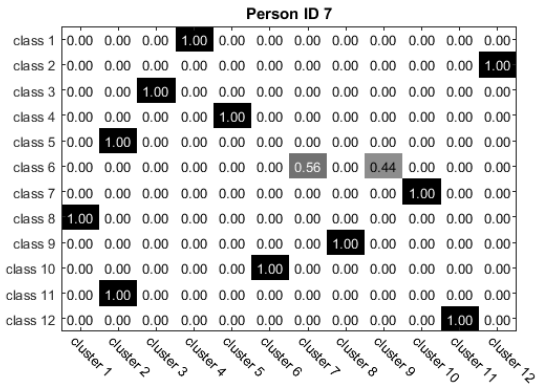

Fig. 10. Confusion matrix after clustering actions into different clusters of subject 7

of performer's body which named "Wind up the music" [11]. On the other hand, actions $class$ 11 were named "Lay down the tempo of a song" presenting by the action of beating the air with both of arms [11]. In the dataset, subject 3 [1] and subject 7 [2] always performed both 2 action classes above in the same way by moving both of the arms in front of their body, as the result, feature vectors are encoded these actions in the similar ways and it was eventually assigned into the same cluster.

According to the $Precision$, $Recall$ and F_{value} receiving from the experiment, the accuracy was acceptable. It was proved that our approach can generally cluster a set of actions into different groups which represent for similar body movements even the original dataset using Kinect sensor was much noisy. The experimental results also convinced that it was an appropriate approach for using low-cost sensors like Kinect for getting information about human body expression during daily human-robot interaction. These information will be clustered into different groups and then robot can utilize the information as a reference to generate its own emotional expression.

5. Conclusion and Future Work

This paper proposed new approach to generate emotional body expression for robot by using human behavior during daily human-robot interaction. We demonstrated our approach by using public dataset for clustering human actions to generate an appropriate representative emotional behaviors. Our future work concerns to automatically generate robot emotional expressions by mapping between representative emotional body expression model and real robot body states. It is believed that robot behavior can also adapt to owner's personal and cultural identity by using the their emotional behaviors for generating robot emotional expression.

[1] represented by dataset P2_1_5_p03 and P2_1_11_p03 for action $class$ 5 and 11 respectively

[2] represented by dataset P2_1_5_p07 and P2_1_11_p07 for action $class$ 5 and 11 respectively

ACKNOWLEDGMENT

This work was supported by the EU-Japan coordinated R&D project on "Culture Aware Robots and Environmental Sensor Systems for Elderly Support (CARESSES)" commissioned by the Ministry of Internal Affairs and Communications of Japan and EC Horizon 2020.

REFERENCES

[1] H. G. Wallbott, "Bodily expression of emotion," *European journal of social psychology*, vol. 28, no. 6, pp. 879–896, 1998.

[2] A. Beck, B. Stevens, K. A. Bard, and L. Cañamero, "Emotional body language displayed by artificial agents," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 1, p. 2, 2012.

[3] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 119–155, 2003.

[4] C. Smith and H. Scott, "A componential approach to the meaning of facial expressions. in russell, ja & fernández-dols, jm (eds.) the psychology of facial expression," 1997.

[5] J. Van den Stock, R. Righart, and B. De Gelder, "Body expressions influence recognition of emotions in the face and voice." *Emotion*, vol. 7, no. 3, p. 487, 2007.

[6] M. Häring, N. Bee, and E. André, "Creation and evaluation of emotion expression with body movement, sound and eye color for humanoid robots," in *Ro-Man, 2011 Ieee*. IEEE, 2011, pp. 204–209.

[7] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *Journal of nonverbal behavior*, vol. 28, no. 2, pp. 117–139, 2004.

[8] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013.

[9] S. Feinman, D. Roberts, K.-F. Hsieh, D. Sawyer, and D. Swanson, "A critical review of social referencing in infancy," in *Social referencing and the social construction of reality in infancy*. Springer, 1992, pp. 15–54.

[10] S. Feinman, "Social referencing in infancy," *Merrill-Palmer Quarterly (1982-)*, pp. 445–470, 1982.

[11] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1737–1746.

[12] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations." in *IJCAI*, vol. 13, 2013, pp. 2466–2472.

[13] J. R. Padilla-López, A. A. Chaaraoui, and F. Flórez-Revuelta, "A discussion on the validation tests employed to compare human action recognition methods using the msr action3d dataset," *arXiv preprint arXiv:1407.7390*, 2014.

[14] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1, pp. 1–6, 1998.

[15] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Transactions on neural networks*, vol. 11, no. 3, pp. 586–600, 2000.

[16] J. Vesanto and M. Sulkava, "Distance matrix based clustering of the self-organizing map," in *International Conference on Artificial Neural Networks*. Springer, 2002, pp. 951–956.

[17] H. Fang, Y. Du, L. Xia, J. Li, J. Zhang, and K. Wang, "A topology-preserving selection and clustering approach to multidimensional biological data," *Omics: a journal of integrative biology*, vol. 15, no. 7-8, pp. 483–494, 2011.

[18] J. Lampinen and E. Oja, "Clustering properties of hierarchical self-organizing maps," in *Mathematical Nonlinear Image Processing*. Springer, 1993, pp. 165–176.

[19] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[20] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation mocap database hdm05," 2007.

[21] D.-D. Nguyen and H.-S. Le, "Kinect gesture recognition: Svm vs. rvm," in *Knowledge and Systems Engineering (KSE), 2015 Seventh International Conference on*. IEEE, 2015, pp. 395–400.