JAIST Repository
https://dspace.jaist.ac.jp/

Title	Incremental Learning of Human Emotional Behavior for Social Robot Emotional Body Expression
Author(s)	Tuyen, Nguyen Tan Viet; Jeong, Sungmoon; Chong, Nak Young
Citation	2018 15th International Conference on Ubiquitous Robots (UR): 377-382
Issue Date	2018-06-27
Type	Conference Paper
Text version	author
URL	http://hdl.handle.net/10119/15482
Rights	This is the author's version of the work. Copyright (C) 2018 IEEE. 2018 15th International Conference on Ubiquitous Robots (UR), 2018, 377-382. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Description	

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Incremental Learning of Human Emotional Behavior for Social Robot Emotional Body Expression

Nguyen Tan Viet Tuyen, Sungmoon Jeong, and Nak Young Chong

***Abstract*— Generating emotional body expressions for social robots has been gaining increased attention to enhance the engagement and empathy in human-robot interaction. In this paper, an enhanced model of robot emotional body expression is proposed which places emphasis on the individual user's cultural traits. Similar to our previous paper, this approach is inspired by social and emotional development of infants interacting with their parents who have a certain cultural background. Social referencing occurs when infants perceive their parents' facial expressions and vocal tones of emotional situations to form their own interpretation. On the other hand, this model replaces the batch learning self-organizing map with the dynamic cell structure, incrementally training a neural network model with a variety of emotional behaviors obtained from the users with whom the robot interacts. We demonstrate the validity of our incremental learning model through a public human action dataset, which will facilitate the acquisition of emotional body expression of socially assistive robots as a reflection of the individual user's culture.**

I. INTRODUCTION

Human facial and bodily expressions play important roles in non-verbal communication to facilitate the recognition of emotions. Psychological researches have shown that emotion and physical expression are an integral part of social interactions to convey how the communicator is feeling and affecting to social outcome [1]. In recent years, many studies focused on generating emotional expressions by estimating and incorporating the emotional states of robot, which is believed to increase the engagement and empathy between humans and robots [2]. In [3], the authors investigated the role of culture in representing robot emotions, where bodily expressions were used as a reliable modality represented for robot emotional state. The study showed that robots can learn to behave socially in alignment to the individual user's cultural traits. The experiment results conveyed that under the effects of cultural differences, robots could generate different emotional and behavioral responses to the same environmental stimuli as shown in Figure 1. Pepper robot's bodily expressions were created in [3] based on the perspectives of social psychology about the connection between emotion and bodily movement [4] [5]. Similarly, emotional expression with bodily movement and eye color for NAO robot was proposed by Markus [6]. This study was mainly motivated by the work of Meijer [7] and other psychological researchers about the contribution of body movements to the attribution of emotions. Likewise, an android head robot developed by Andra [8] imitates human facial expressions with the main goal to improve the emotion recognition capabilities of autistic children. The android robot tracks human expression represented by facial feature points and directly convert them into corresponding motor movements of robot.

The authors are with the School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan {ngtvtuyen, jeongsm, nakyoung}@jaist.ac.jp

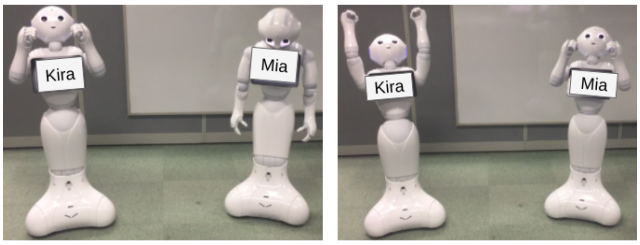

Fig. 1. Pepper robot conveys its emotion though bodily expressions

On the other hand, in order to increase the engagement of the conversation and the empathy between a robot and a human through long term interaction, careful attention should be paid to generate robot emotional expressions according to the personality and cultural identity of a person. This assumption has been strongly supported by straightforward relation between individual cultural traits and robot behaviors which has been found by HRI researcher [9] or psychological evidence about mimicry of posture, facial expression, verbal and non-verbal behaviors of interacting partners [10]. To archive this goal, this research was motivated by the psychological perspectives about infant social development where the infant's interpretation and behaviors are highly influenced by their parents through imitative exchanges [11]. Infant is rapidly influenced by the guideline from their parents in acquiring knowledge about typical event. They generate emotion and behavior in response to the stimuli by the imitative mechanism to form their own emotions and behaviors as similar as encoded emotion and expression from their parents. An interesting example was mentioned in [12] 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 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 bodily expressions of socially assistive robots, allowing robots to enter into natural and intuitive social human robot interactions. Motivated from that, this paper propose that, during social interactions, the robot should pay attention to their owner's emotional bodily expressions associated with a typical emotional state as

its interested stimuli. Through long term interaction, robot generates emotional bodily 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 human habitual behaviors which could be measured by assessing the frequency of past behaviors [13] as the reference for generating the robot's emotional bodily 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 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 expressions as reference information. In the Experiments and Results section, a public data set was used to evaluate our proposed model. The discussion about experimental results and future work are mentioned in the Discussion part. Finally, we summarize the results as well as our future work in the Conclusion and Future Work section.

II. Methodology

During daily human robot interactions, it is obvious that, for each typical emotion, the number and types of human emotional gestures vary depending on the user's cultural and personal identity, which are unknown beforehand. Thus, robots are required to be capable of learning undefined behaviors in an unsupervised manner. This idea has been shared across many previous researches, Mohammad [14] used unsupervised learning for association between human gestural commands and robot actions. In [15], the authors made comparisons between different unsupervised learning algorithms such as Self Organizing Maps (SOM), Fuzzy C Means (FCM) and K Means for recognizing human postures in video sequences. The capability of trajectory learning from human demonstrations for the robot arm was proposed by [16], where the trajectory clustering and approximation modules take human demonstrative trajectories as the input that belong to different clusters. For each group, the most consistent trajectory has been selected and then a set of generated trajectories can be visualized in simulated environment, thereby human finally can select the desired trajectory. Hence, previous studies through this section convinced that for scenarios of interaction, while prior information about actions are not available, the unsupervised learning is an appropriate approach for classifying body movements into different groups based on the similarity of actions.

In order to obtain human bodily expression information, for each typical emotion, motion capture sensors like Kinect can be easily used to extract demonstrator's skeleton. A set of human emotional behavior $A_1, A_2, ..., A_n$ are gradually received during day-to-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. The Covariance Descriptor method [17] is used to

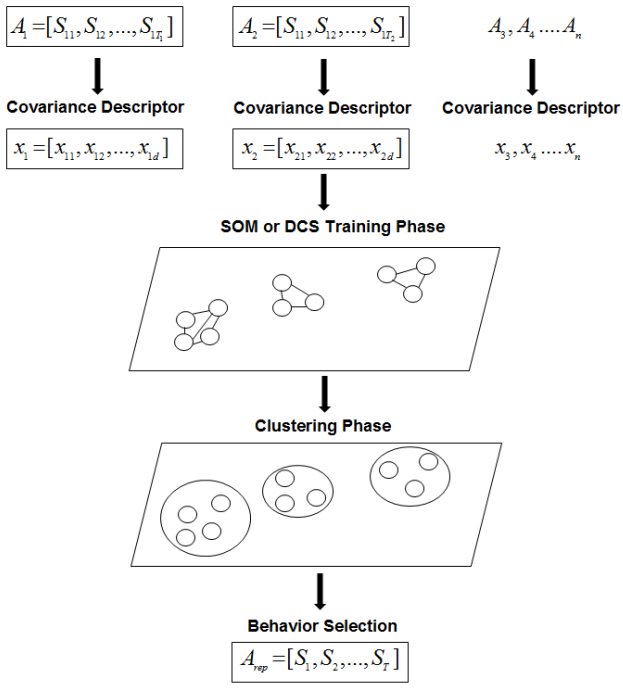

Fig. 2. The model of behavior selection through long term interaction

encode the sequence of frames A_i into a fixed length descriptor. Human emotional body expression $A_1, A_2, ..., A_n$ are classified into j clusters through training and clustering phase. Finally, considering the distribution of body movements, robot can utilize the most frequently observed behavior as the reference for generating its emotional bodily expression. Figure 2 presents our model of behavior selection to generate robot emotional expression on typical emotion space such as *Happy*.

A. Training Phase

1) Self Organizing Map: In our previous paper [18], Self Organizing Map (SOM) [19] was utilized as the batch learning approach for the training phase in behavior selection as shown in Figure 2. SOM was conducted as unsupervised learning using no prior knowledge about number of clusters which is suitable for the scenarios of daily interactions as discussed before. SOM ensures the topological properties of the descriptors were preserved after reducing from the high-dimensional input space to the low-dimensional space grid. Meaning that, if two different behavior samples were closed to each other on the original feature space, they should be remained with similar topological properties in different dimensional grid.

2) Dynamic Cell Structure: It is obvious that topological preservation is the main advantage of SOM for classifying encoded descriptors into different groups based on the similarities. On the other hand, during long term human robot interaction, number of human emotional behaviors will be sequentially increased. Thus, the robot needs to be capable of incrementally learning new gestures without corrupting the previous model. To satisfy such requirements, Dynamic

Cell Structure (DCS) [20] is an appropriate approach where topological properties could be preserved in a similar way to SOM. Indeed, DCS provides capability of learning new behavior samples as the incremental manner.

DCS represents family of artificial neural networks which could be applied for both supervised and unsupervised manner. It belongs to class of Topology Representing Networks which build perfectly topology preserving feature maps [21]. DCS inheres Kohonen type learning rule [19] for updating weight of neural vectors as SOM while using Hebbian learning rule [22] to dynamically update lateral connection structure (topology of the graph of neurals). Another approaches of growing neural network by dynamic allocation the feature map in order to evolve its structure are known as Growing Cell Structure (GCS) [23], Growing Gas Model or Growing Neural Gas [24]. DCS works in a similar way to GCS excepts one essential difference: the lateral connections between neural units are not initially defined, instead, they are dynamically learned during training phase [20] by Herbian learning rule. DCS has been widely used in many applications for on-line learning purpose, NASA's first generation Intelligent Flight Control System program utilized DCS for on-line learning and estimation of system parameters [25].

The unsupervised DCS starts with initializing 2 neural units m_1 and m_2, they are connected to each other by lateral connection of weight $C_{12} = C_{21} = 1$. It is noted that lateral connection of neurons is defined in the range from 0 to 1. $C_{ij} = 1$ if they are completely connected to each other and vice versa, $C_{ij} = 0$ if they are disconnected to each other. Lateral connections in DCS are always bidirectional and have symmetric weights.

For the input descriptor x_i, neuron located nearest m_{bmu} and second nearest m_{second} to descriptor x_i are firstly determined by Equation (1). The neighboring neurons of m_{bmu}, represented by N_{bmu}, are updated lateral connections by Herbian learning rule [22]. This rule is mathematically described as Equation (2) whereas ε is forgetting constant, θ is threshold for deleting lateral connection.

$$\begin{aligned} ||x_i - m_{bmu}|| \leq ||x_i - m_i||, 1 \leq i \leq N \\ ||x_i - m_{second}|| \leq ||x_i - m_i||, 1 \leq i \neq bmu \leq N \end{aligned} \tag{1}$$

$$C_{ij}(t+1) = \begin{cases} 1 & : (i = bmu) \wedge (j = second) \\ 0 & : (i = bmu) \wedge (j \in \{N_i\}, \\ & \quad j \neq second) \wedge (C_{ij} < \theta) \\ \varepsilon C_{ij}(t) & : (i = bmu) \wedge (j \in \{N_i\}, \\ & \quad j \neq second) \wedge (C_{ij} \geq \theta) \\ C_{ij}(t) & : otherwise \end{cases} \tag{2}$$

DCS then updates the weight of neuron vectors by Kohonen learning rule [19] which makes them move closer to the current input:

$$\begin{aligned} m_{bmu} = m_{bmu} + \alpha(t) \times (x_i - m_{bmu}) \\ m_i = m_i + \alpha(t) \times \phi(m_i, m_{bmu}) \times (x_i - m_i), \end{aligned} \tag{3}$$

where m_i is neighbors of neuron m_{bmu}, α is the learning rate and $\phi(m_i, m_{bmu})$ is the neighborhood kernel function.

A training cycle is finished with updating resource value of best matching unit neurons as Equation (4). If quantization error did not drop under the stopping condition (the predefined accuracy), the new neuron unit m_{new} should be inserted into the network and located between neurons with largest and second largest resource value. Finally resource of all neurons unit will be decreased as Equation (5) with β is a decay constant ($0 \leq \beta \leq 1$).

$$\Delta\tau_{bmu} = ||x_i - m_{bmu}||^2 \tag{4}$$

$$\tau_i(t+1) = \beta\tau_i(t) \tag{5}$$

Similar to SOM, DCS ensures the topological property on grid of trained neurons. Indeed, DCS dynamically modifies the lateral connections by Herbian learning rule [22] and adding new unit on the grid of neurons if the quantization error is still higher than stopping condition. After grid of neurons had been trained, these neurons will be classified into different groups at the clustering phase.

B. Clustering Phase

At the clustering phase, classifying trained neurons into different groups is conducted with Distance matrix based approach [26]. By clustering the training neurons rather than descriptors directly, significant gains in speed of clustering can be obtained [27]. At the end of the clustering phase, each neuron and its corresponding descriptor was defined by Best Matching Unit (BMU) function:

$$||x - m_i|| = min\{||x - m||\} \tag{6}$$

C. Behavior Selection Phase

Until this step, n actions $\{A_1, A_2, ..., A_n\}$ were encoded to n descriptors $\{x_1, x_2, ..., x_n\}$ and then classified into different groups $\{Cluster_1, Cluster_2, ..., Cluster_k\}$ ($k \leq N$) based on the similarity of action movement. At the behavior selection phase, by considering the distribution of action observations, an appropriate behavior will be selected. Here, the most distributed cluster $Cluster_i$ ($i \in k$) included the highest number of similar actions, as the result, representative descriptor x_{rep} located nearest to the center of this cluster is defined as:

$$||x_{rep} - center|| \leq ||x - center|| \quad \forall x \in Cluster_i, \tag{7}$$

where $||x - center||$ is the Euclidean distance between center of $Cluster_i$ to descriptor x . Finally, the corresponding action of descriptor x_{rep} will be detected as A_{rep}. Robot can utilize human habitual action A_{rep} as reference to generate its emotional bodily expression associated with corresponding emotion.

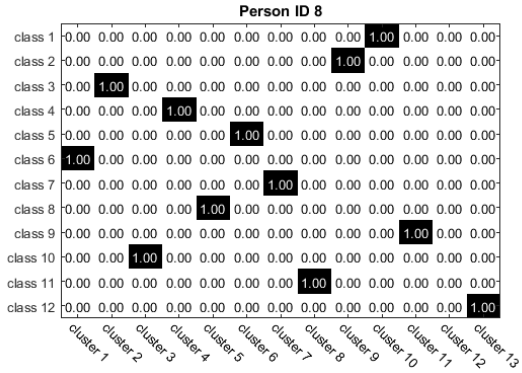

Fig. 3. Confusion matrix of subject 8 representing for percentage of actions belong to $class\ i$ were assigned into $cluster\ j$ conducted by SOM

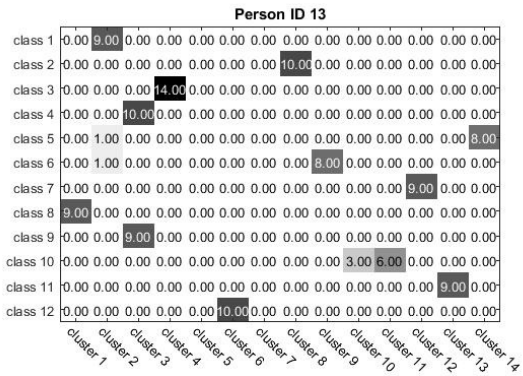

Fig. 4. Confusion matrix representing for number of actions belong to $class\ i$ were assigned into $cluster\ j$ from subject 13 conducted by DCS

III. Experiment and Results

To validate performance of SOM and DCS training phase in behavior selection model, experiment setup was conducted in the similar way to our previous paper [18] where the public Microsoft Research Cambridge-12 Kinect gesture dataset (MSRC-12) [28] was utilized. It was assumed that these gestures were acquired from human emotional expression during daily human robot interaction in the same emotion space as *Happy*. $Precision$, $Recall$ and F_{value} were used as evaluation criteria for this experiment. It was noticed that no prior information about action classes (true labels) or number of classes had been used, meaning that this experiment aimed to measure performance of unsupervised learning SOM and DCS for clustering original actions into different groups based on the similarity of body movements.

A. Experiment Results

Each subject data were encoded into the corresponding descriptor. Then, a set of covariance descriptors was used for training a gird of neurons as training phase. On the grid of trained neurons, a set of local neurons were firstly determined from clustering phase. Then, the rest of neurons were assigned into appropriate clusters by minimizing the distance between determined local representative neurons and them. Since topological properties of the feature descriptors were preserved on the grid of SOM (and DCS) 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 can be defined by BMU Equation (6), meaning that feature descriptors were assigned into the same clusters if its corresponding neurons located in the same clusters. Figure 3 presents the example of clustering result by using the 8-th subject's data.

Experiment procedures were repeated by replacing with DCS training phase where subject data was incrementally input to the proposed model. To this end, in order to evaluate the proposed model using entire dataset, we calculated the average values of $Precision$, $Recall$ and F_{value} using 15 subjects data as shown in Table I.

TABLE I

The average values after 15 experiments conducted by SOM

	SOM	DCS
$Precision$	0.9166	0.8019
$Recall$	0.9115	0.9141
F_{value}	0.9133	0.8524

IV. Discussion

To analyze the proposed model in more detail, we use the 13-th person's data which was conducted by DCS as shown in Figure 4. It was obvious that among 14 different clusters, $cluster\ 4$ was the most populated cluster with 14 similar actions $A_1, A_2, ..., A_{14}$ located inside. The most populated cluster means the most frequent human body expression for *Happy* which robot observed during social interactions. As the result, an action $A_i \in \{A_1, A_2, ..., A_{14}\}$ which was defined by Equation (7) was selected as the representative action to generate robot bodily expression *Happy*. 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.

Neurons and its corresponding feature descriptors were defined by BMU Equation (6). However, it was noticed that, the number of clusters on the grid of trained 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. In Figure 5, there were 13 clusters created by a grid of SOM neurons representing for action data set of subject 8. The corresponding confusion matrix in

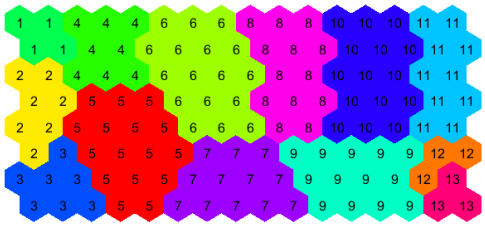

Fig. 5. 13 clusters were detected from grip of SOM neurons which represent for action dataset of subject 8

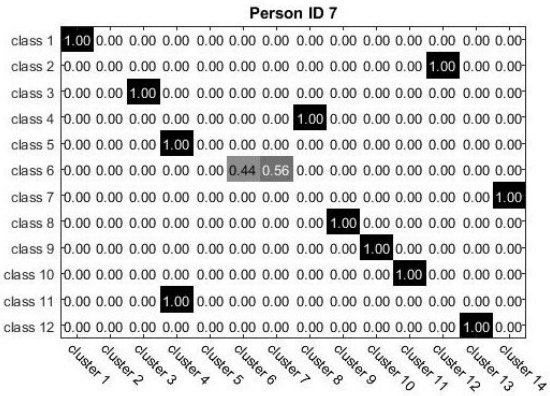

Fig. 6. Confusion matrix of subject 7 conducted by DCS training phase

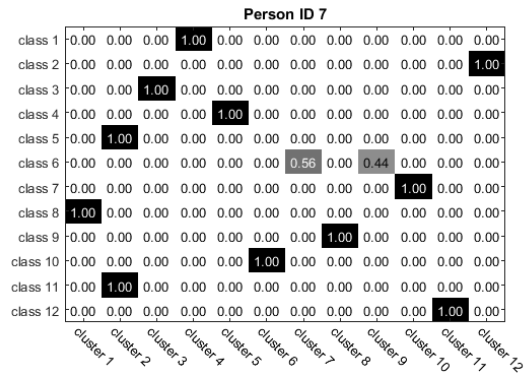

Fig. 7. Confusion matrix of subject 7 conducted by SOM training phase

Figure 3 indicates that there were actually 12 clusters created since there were no elements located in *cluster* 12. This problem also occurred in DCS model as shown in Figure 6 since no action was available at *cluster* 2 and *cluster* 5. In other words, only 12 clusters were properly created.

Because MSRC-12 dataset, which was gathered by Kinect sensor, was much more noisy than HDM05-MoCap dataset [29] obtained from optical marker sensors [30], as the result, same action class can be divided into many sub-clusters. The experiment carried out with SOM noticed that in subject 7, actions *class* 6 were divided into *cluster* 7 and *cluster* 9 respectively as shown in Figure 7 . Similarly, experiment conducted by DCS with that subject as presented in Figure 6 reveals that actions *class* 6 were separated into *cluster* 6 and *cluster* 7.

MSRC-12 dataset includes 12 different actions classes.

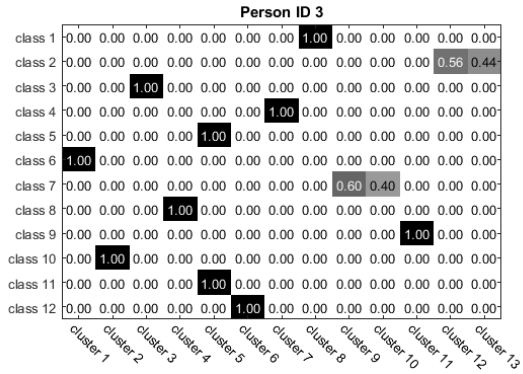

Fig. 8. Confusion matrix of subject 3 conducted by SOM training phase

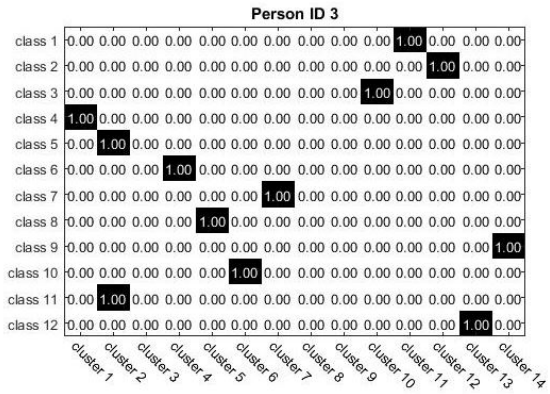

Fig. 9. Confusion matrix of subject 3 conducted by DCS training phase

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, SOM clustered 100% actions *class* 5 and 100% actions *class* 11 into the same *cluster* 5 as shown in Figure 8. Consequentially, actions *class* 5 and *class* 11 of subject 3 were also located into the same *cluster* 2 as shown in Figure 9 when DCS was applied to this dataset. In terms of subject 7, both action *class* 5 and *class* 11 were located in the same *cluster* 2 as presented in Figure 7 when conducted with SOM. Similarly, they belong to the same *cluster* 4 when experimented with DCS as shown in Figure 6. Actions *class* 5 were described as the movement of both arms in front of the user's body which named "Wind up the music" [28]. 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 [28]. 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 encoded these actions in similar ways and it was eventually assigned into the same cluster.

[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

According to the $Precision$, $Recall$ and F_{value} received from the experiment, it was clear that the accuracy in both of SOM and DCS were generally competitive. In other words, unsupervised learning approach can successfully cluster a set of actions into different groups which represent for similar body movements even the original dataset using Kinect sensor was much noisy. Thus, the low-cost motion capture sensor like Kinect was an appropriate approach for obtaining information of human body expression during daily human robot interaction. Secondly, the evaluation results proved that unsupervised batch learning like SOM showed better performance than the other one. However, that accuracy of DCS was acceptable whereas incremental learning DCS gained considerable benefit on time processing data and computation cost compared to SOM batch learning. These factors play crucial roles in social human robot interaction scenarios while the amount of information was sequentially increased and robot is required capability of incrementally updating model by learning from new stimuli without removing previous one. Overall, DCS should be considered as an unsupervised learning manner in the proposed model for learning human emotional expression in social interactions.

V. Conclusion and Future Work

This paper addressed an incremental learning algorithm of human emotional behaviors toward generating emotional body expressions for social robots through human robot interactions. The proposed idea was demonstrated with a public dataset for clustering human actions and then generating appropriate representative behaviors. Our future work will focus on efficient imitation models to transfer the obtained representative behavior into a variety of robot kinematic models. It is believed that social robot gestures can be better adapted to the individual user's cultural traits through a prolonged period of interactions.

Acknowledgments

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

References

[1] B. N. Vosk, R. Forehand, and R. Figueroa, "Perception of emotions by accepted and rejected children," *Journal of Psychopathology and Behavioral Assessment*, vol. 5, no. 2, pp. 151–160, 1983.

[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] T. L. Q. Dang, N. T. V. Tuyen, S. Jeong, and N. Y. Chong, "Encoding cultures in robot emotion representation," in *IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2017.

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

[5] 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.

[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. De Meijer, "The contribution of general features of body movement to the attribution of emotions," *Journal of Nonverbal behavior*, vol. 13, no. 4, pp. 247–268, 1989.

[8] A. Adams and P. Robinson, "An android head for social-emotional intervention for children with autism spectrum conditions," in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 183–190.

[9] E. Park, D. Jin, and A. P. del Pobil, "The law of attraction in human-robot interaction," *International Journal of Advanced Robotic Systems*, vol. 9, no. 2, p. 35, 2012.

[10] T. L. Chartrand and J. A. Bargh, "The chameleon effect: the perception–behavior link and social interaction." *Journal of personality and social psychology*, vol. 76, no. 6, p. 893, 1999.

[11] S. Feinman and M. Lewis, "Social referencing at ten months: A second-order effect on infants' responses to strangers," *Child development*, pp. 878–887, 1983.

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

[13] I. Ajzen, "Residual effects of past on later behavior: Habituation and reasoned action perspectives," *Personality and social psychology review*, vol. 6, no. 2, pp. 107–122, 2002.

[14] Y. Mohammad, T. Nishida, and S. Okada, "Unsupervised simultaneous learning of gestures, actions and their associations for human-robot interaction," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE, 2009, pp. 2537–2544.

[15] K. K. Htike and O. O. Khalifa, "Comparison of supervised and unsupervised learning classifiers for human posture recognition," in *Computer and Communication Engineering (ICCCE), 2010 International Conference on*. IEEE, 2010, pp. 1–6.

[16] J. Aleotti and S. Caselli, "Robust trajectory learning and approximation for robot programming by demonstration," *Robotics and Autonomous Systems*, vol. 54, no. 5, pp. 409–413, 2006.

[17] 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.

[18] N. T. V. Tuyen, S. Jeong, and N. Y. Chong, "Learning human behavior for emotional body expression in socially assistive robotics," in *Ubiquitous Robots and Ambient Intelligence (URAI), 2017 14th International Conference on*. IEEE, 2017, pp. 45–50.

[19] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[20] J. Bruske and G. Sommer, "Dynamic cell structure learns perfectly topology preserving map," *Neural computation*, vol. 7, no. 4, pp. 845–865, 1995.

[21] I. Ahrns, J. Bruske, and G. Sommer, "On-line learning with dynamic cell structures," in *Proceedings of the International Conference on Artificial Neural Networks*, vol. 2, 1995, pp. 141–146.

[22] T. Martinetz, "Competitive hebbian learning rule forms perfectly topology preserving maps," in *ICANN93*. Springer, 1993, pp. 427–434.

[23] B. Fritzke, "Growing cell structuresa self-organizing network for unsupervised and supervised learning," *Neural networks*, vol. 7, no. 9, pp. 1441–1460, 1994.

[24] Fritzke, "A growing neural gas network learns topologies," in *Advances in neural information processing systems*, 1995, pp. 625–632.

[25] M. G. Perhinschi, G. Campa, M. R. Napolitano, M. Lando, L. Massotti, and M. L. Fravolini, "A simulation tool for on-line real time parameter identification," in *Proceedings of the 2002 AIAA modeling and simulation conference*, 2002.

[26] 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.

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

[28] 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.

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

[30] 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.