Affect State Recognition for Adaptive Human Robot Interaction in Learning Environments

Dimitrios Antonaras, Charis Pavlidis, Nicholas Vretos and Petros Daras
Centre for Research and Technology Hellas, Information Technologies Institute,
Thessaloniki, Greece.
Email: {dantonar, pavlidis, vretos, daras}@iti.gr

***Abstract*—Previous studies of robots used in learning environments suggest that the interaction between learner and robot is able to enhance the learning procedure towards a better engagement of the learner. Moreover, intelligent robots can also adapt their behavior during a learning process according to certain criteria resulting in increasing cognitive learning gains. Motivated by these results, we propose a novel Human Robot Interaction framework where the robot adjusts its behavior to the affect state of the learner. Our framework uses the theory of flow to label different affect states (i.e., engagement, boredom and frustration) and adapt the robot's actions. Based on the automatic recognition of these states, through visual cues, our method adapt the learning actions taking place at this moment and performed by the robot. This results in keeping the learner at most times engaged in the learning process. In order to recognizing the affect state of the user a two step approach is followed. Initially we recognize the facial expressions of the learner and therefore we map these to an affect state. Our algorithm perform well even in situations where the environment is noisy due to the presence of more than one person and/or situations where the face is partially occluded.**

I. Introduction

Humanoid robots that have been used in learning environments show that the interaction with a robot could bring about cognitive learning gains ([2], [10]). However, the interaction with the robot could be either adaptive or non-adaptive during the teaching procedure. An adaptive interaction is where the robot adjusts its behavior according to the learning needs or the state of a user, whereas, non-adaptive is the case where the robot behaves equally for every user regardless its cognitive state. For instance, in [5] the robot was giving instructions about a puzzle game according to learner's progress in the game and in [4] the robot was interacting with a user according to its affect state and body posture. On the other hand, in [2], [3], [10] the robots were giving predefined instructions without any adaptation. Comparing the two cases, research in the field ([4]-[9]) shows that the adaptive interaction outperforms the non-adaptive resulting to more cognitive learning gains.

This work, proposes a framework for Human Robot Interaction in learning environments, where the robot attempts to recognize the affect state of the learner during the teaching procedure. Afterwards, the robot is able to adapt the learning procedure according to the affect state of the learner aiming to keep her/him engaged. For instance, in the case where the robot plays a learning game with the user and the user is starting to feel frustration. Provided that the robot has this information, it could change the difficulty of the game or the entire game in order to engage the learner to the learning process. Regarding the affect state of the learner we first analyze her/his facial expressions using the robot's camera and subsequently we assign it to an affect state as in [11].

978-1-5386-0756-5/17/$31.00 ©2017 IEEE

The rest of the paper is organized as follows. Section II, includes related works for robots in learning environments and affect recognition. Our framework and its setup will be analyzed in detail at section III. Finally, section IV contains the results of our experiments and section V a sum up of our work.

II. Related Work

In [1] the authors present an extended review with respect to the deployment of robots in education, wherein different aspects are presented such as the subject of the learning activity, the place where the deployment took place (during or after the school hours), the role of the robot (tutor, peer, tool), the design of the robot (e.g., low-cost, humanoid) and some theories supporting the usage of robots. The humanoid robot Nao has been used in [3] as an assistive tool for teaching English language in junior high school students where the learning procedure has been adapted to the learner's level. Their experiment consist of two classes, with and without robot appearance, showing that the usage of a robot in a class could improve the learning experience. Kai-Yi Chin et al. in [2] have proposed a Robot-based learning system for applying robots in the elementary education. The experiments took place in two classes of a Taiwanese school where they applied their system in one of them whereas in the other one they used a PowerPoint-based learning system. The robot has been used as a tutor giving instructions for the subject, while it was supervised by a teacher. In the PowerPoint-based system the instructions has been given through loudspeakers synchronized with PowerPoint slides. Finally, the experiments have been evaluated in a post-test, pre-test scheme, showing that the former classroom outperformed the latter, concluding that the usage of a robot in a classroom could have positive effects in the learners performance. Kapoor and Picard in [4] are proposing a framework for affect recognition in learning environments where information from multiple sensors is combined to recognize the affect state of a pupil. In details, pupils are playing a game while they are monitored from an IR camera, to extract upper and lower face features and a pressure sensing chair, to extract body postures features. Regarding the affect state recognition algorithm, it uses machine learning techniques to combine the features and the current state of the game in a multimodal fusion scheme. Finally, the algorithm has been

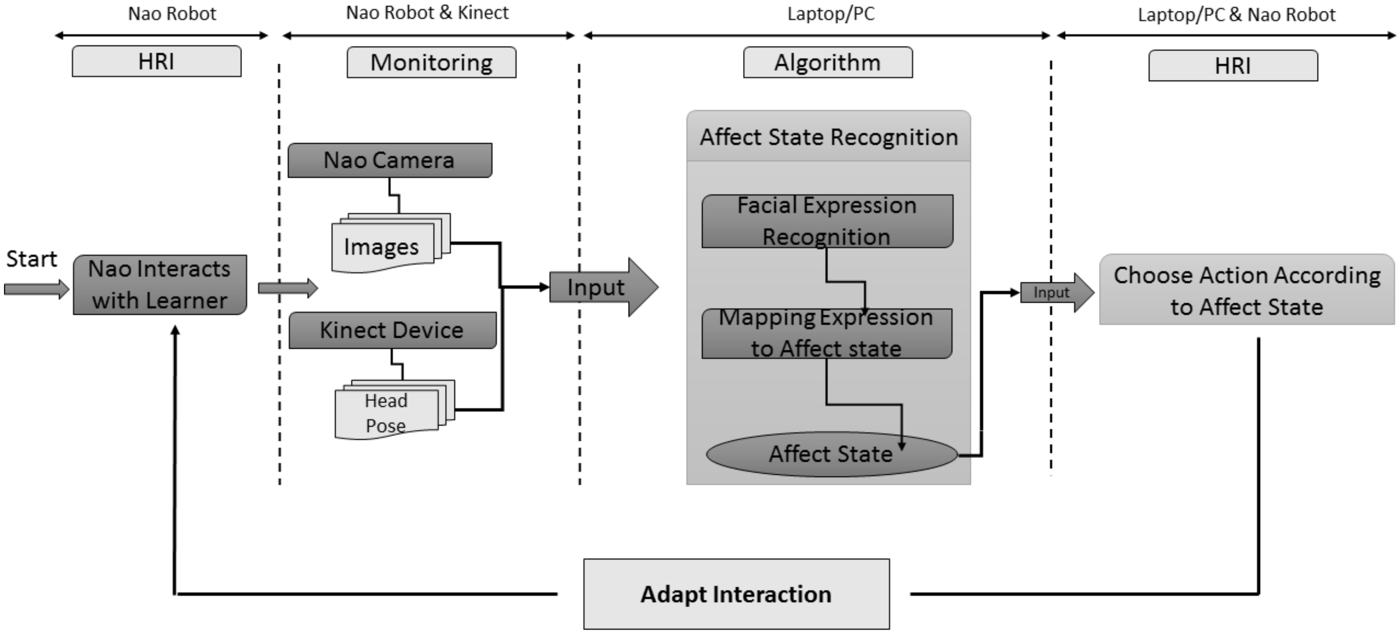

Fig. 1. A full cycle of our framework. Nao robot begins interaction with the learner, e.g., a learning task. Robot's camera and Kinect device inputs the affect recognition algorithm with images of the learner and her/his head pose respectively. The algorithm outputs the affect state of the learner and the robot decides for its next action accordingly.

trained on a database created by the same group. Leyzberg et al. have already shown in [9] that humanoid robots can have a positive effect on the learning experience, whereas in [5] they study the case where the robot is personalized according to the learners' needs. The learners have been asked to solve a series of puzzle games, while a robot has been placed aside giving personalized or non-personalized directions (relative to the puzzle) helping them to find the solution. To personalize the directions, the robot was taking feedback from the progression of the game. Finally, they compare the two cases by measuring the time spent for solving the puzzles, showing that the personalized case outperforms the non-personalized. The authors in [10] have tried to investigate the impact of social and asocial robots in education. Comparing the two robots, the asocial robot have had less sentences for verbal communication, its gestures were not synchronized with the context of the speech, there was no personalization feature (e.g., calling the name of the pupil) and its gaze were neither on the pupil or the game. The educational context was to learn the prime numbers in children of age 7 and 8. They concluded that robots could bring about cognitive learning gains comparing to traditional teaching procedures, however, comparing the social and asocial robot, the results show that the latter outperformed the former.

The affect detection from facial features is a well-researched area and numerous studies have been published. Considering systems that attempt to detect learning-centered affective states, an initial study is provided in [4]. Hoque et al. [12] try to classify smiles as either frustrated or delighted. The authors extract facial features from temporal information of video and therefore a classifier was used to accurately distinguish between frustrated and delighted smiles correctly in 92% of the cases. Next, Grafsgaard et al. [13] used the Computer Expression Recognition Toolbox (CERT) [14], which is a computer vision tool used to automatically detect Action Units (AUs). The particular tools aims to recognize the level of frustration and cognitive. The authors correlate the presence of specific AUs with frustration and cognitive gain. Whitehill et al. [15] create automated engagement detectors distinguishing between high and low engagement. By extracting appearance-based features using Gabor filters and by using a support vector machine, they manage to achieve the level of engagement. Bosch et al. [16] use computer vision and machine learning techniques to detect students affect (boredom, confusion, delight, engagement and frustration). Students play an education game while facial expressions and gross body movements are gathered. They use CERT to extract facial features (AUs, orientation and position of the face). The researchers established classification models for seven discriminations (overall, five affective state models, and off task vs. on task) using 14 different classifiers to test models performance.

III. Proposed Method

A. Overview

The setup of our experiments consists of the humanoid robot Nao, a Kinect v2 sensor and a laptop/PC, where the affect state recognition algorithm resides. The input of our algorithm is the images from Nao's camera and the user's head pose from the Kinect device. The head pose of the user is used as a flag for our algorithm to determine when the user is facing the robot and thus start processing the images. Finally, the robot is interacting with the user and simultaneously takes feedback from the algorithm to adapt its actions according to the affect of the learner.

Note that, the head pose of the user returned from the Kinect device is a 3D vector where its coordinates are related to the Kinect sensor, therefore, a transformation from Kinect to Nao referential system is required. To achieve this, we calibrate the two cameras so as to obtain the relative position of the two cameras in the space and subsequently transform the vector.

B. Kinect-Nao Calibration

The calibration procedure follows a standard stereo camera calibration scheme consisting of a predefine pattern (checkerboard), which help us to find matching points between the 2 cameras. Having these correspondences we can therefore calculate the transformation between the two cameras, and thus, the two devices.

We first capture images of a checkerboard in different postures (sort rotations in x, y, z axes) where both cameras can view it and then using the Stereo Calibration Tool of Matlab we obtain the translation vector $\mathbf{t}$ and rotation matrix $\mathbf{R}$ of the two cameras. Then, we apply them to the head pose vector $\mathbf{v}$:

$$\mathbf{v}_{NaoWorld} = \mathbf{v}_{KinectWorld}\mathbf{R} + \mathbf{t} \tag{1}$$

C. Affect State Recognition

This approach takes advantage of the ability of face representation as a graph. The face is located using points tracing specific areas of the face, which are then used to create a graph. The variation of muscle movement on the face, during the expression of different emotions, leads to different positions of points on the image and generates different graphs. The algorithm uses this graph variation to predict the different emotions.

The algorithm takes as input an image, then it detects facial landmarks using the Supervised Descent Method (SDM) [17] technique. These landmarks are used for feature extraction, wherein a pre-trained classifier takes into account the extracted features to make a decision about the emotion portrayed in the given image.

The classifier used throughout the whole process is a Support Vector Machine (SVM) a widely used multi-class classifier. Multi-class SVMs classify test datum instances (features) into one of multiple pre-defined target classes, choosing the class that classifies an instance with the greatest margin from other classes.

1) Feature extraction: Facial landmarks are points on specific part of the facial image, which for instance indicate the location of the nose, the eyes, the brows and the mouth within an image. These points are tracked to follow the facial muscles movements in time. Assuming that all facial landmarks are considered as a connected graph, we accept that the density of the graph differs in each facial expression (e.g., the density of connected landmarks around different areas of the graph differs due to an emotion response, differently for each of the examined emotions). Graphs are highly useful mathematical tool that can provide a wealth of information regarding the interrelationships of spatial points in the particular case, of the facial landmarks. In order to extract features from these facial landmarks, spectral graph analysis is implemented, through which a characteristic vector, depicting areas of density in a graph, is extracted. To do so, the Laplacian matrix of the graph is calculated (cf. Formula 3) and the eigendecomposition problem for the eigenvectors corresponding to the 1st, 2nd and 3rd greatest eigenvalues is solved. This eigenvector holds information regarding the different density areas of the initial graph. In this case, these areas are the characteristic areas of eyes, mouth and nose, thus, the areas that are more expressive when an emotion response is triggered.

Given a graph, its combinatorial Laplacian matrix can be defined as:

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \tag{2}$$

Where $\mathbf{D}$ is the degree matrix defined as $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$, where $\mathbf{A}_{ij}$ the elements of $\mathbf{A}$ and $\mathbf{A}$ is the adjacency matrix of the graph computed as:

$$\mathbf{A}_{ij} = 1 - e^{\frac{(-||x_i - x_j||)}{d}} \tag{3}$$

$||x_i - x_j||$ is the Euclidean distance between landmark points, x_i, x_j where $x_i = (a, b)$ is a landmark point on the image grid. d is a constant depicting the variance of the overall distance between the facial landmarks.

In order to normalize between different facial image scales and sizes, a robust version of the Laplacian matrix is used, the so-called symmetric normalized Laplacian matrix which can be calculated as:

$$\mathbf{L}^{Sym} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{(-\frac{1}{2})} \tag{4}$$

Once the symmetric normalized Laplacian matrix is calculated, its eigen-decomposition is considered.

$$\mathbf{L}^{Sym}\mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{5}$$

The corresponding eigenvectors of first, second and third largest eigenvalues are used as the feature of a specific frame.

2) Facial expressions classification: Support Vector Machines (SVM) are one of the most popular supervised learning models for classification that are used in machine learning. The proposed method uses SVM as a classifier, using the RBF kernel function to firstly undergo a training procedure, where labelled ground truth data are used in order to train the algorithm to classify pre-defined labels, based on the association of labelled data with features (in this case facial landmarks) within the training set. In order to train our classifier, a publicly available database was used, i.e., the Cohn-Kanade database [17]. This dataset is limited to labelling images with the well-known 6 spontaneous emotions of Ekman [18], as there is no existing dataset trained after time-dependent affective states, such as engagement, boredom, etc. The Eckmanian emotions are Anger, Disgust, Fear, Happiness, Sadness and Surprise. These emotions, according to Ekman, are the most basic emotions that can be expressed through facial expressions. Baker et al. [11] map the learning-centered cognitive affective states on Russells core-affect framework (2003). In this framework valence (pleasure to displeasure) and arousal (activation to deactivation) compose an affective state. So Boredom has a negative valence and low level of arousal, Frustration has a high negative valence and a high level of arousal and Engaged concentration has a positive valence. The affect states and basic Ekman emotions are represented as points. So, we consider a correlation between the adjacent points, allowing us to direct map the spontaneous emotions to affect states (Table I).

TABLE I. EMOTION WITH AFFECT STATES MAPPING ASSOCIATIONS

Ekmanian Emotion	Time-dependent affect state
Sad	Boredom
Happy	Engagement
Surprise, Anger, Fear	Frustration

Through these associations our algorithm was trained in order to predict affective states related to the theory of flow (boredom, frustration and engagement).

IV. Experimental Results

Using peaked[1] images from the Cohn-Kanade database to train the SVM classifier in the method described in the previous sections (around 80% of all peaked images were used as the training set) and the remaining 20% of peaked images as the test set, the proposed approach yielded the results seen in Table II (representing the values in the confusion matrixs diagonal). We test the performance of our algorithm using several setups. Firstly, we test the performance of our approach predicting 6 label one for each of the basic emotions. The correlation described above in order our algorithm to predict the flow theory states, can used either as a separated labels (Sad for boredom, happy for Engagement and surprise, anger, fear for frustration) to train our SVM or the map of flow theory state from Ekman labels is done after the SVM prediction 6 Ekman emotions.

TABLE II. PERFORMANCE OF OUR APPROACH USING CK DATABASE

Classes	**Accuracy**	**Accuracy (without data correlated to disgust emotion)**
Ekman emotions	0.9167	-
Theory flow states map before classifier prediction	0.9417	0.9490
Theory flow states map after classifier prediction	0.9083	0.9184

Data correlated to disgust emotion in Cohn-Kanade database can be removed due to the fact that this emotion does not map to any of the Theory of Flow states. A confusion matrix was used to evaluate the performance of the classifier over the test set, resulting to an overall accuracy (calculated as the sum of the diagonal of the matrix divided by the entire matrixs sum) of 94,9% for the best setup.

V. Conclusion

Aiming to achieve cognitive gains, we propose a framework for deploying robots in learning environments where robot monitors the teaching procedure to adapt the interaction. As a criterion for the adaptation task, we attempt to recognize the affect state of the learner and subsequently we adjust robot's actions accordingly. Initially, the facial expressions of the learner are being recognized by the affect state recognition algorithm and as a following step it maps the expressions to affect states. Considering future research, testing our framework beyond laboratory conditions, e.g., real learning environments, is an issue that needs to be examined. The online re-calibration of the two cameras after a robot's movement is an additional issue that need to be considered, while the existent calibration is constrained to stationary cameras.

Acknowledgements

The work presented in this document is a result of MaTHiSiS project. This project has received funding from the European Union's Horizon 2020 Programme (H2020-ICT-2015) under Grant Agreement No. 687772.

References

[1] O. Mubin, CJ. Stevens, S. Shahid, A. Al Mahmud and JJ. Dong, *A review of the applicability of robots in education. Journal of Technology in Education and Learning*, Journal of Technology in Education and Learning 1, 2013

[2] K.Y. Chin, Z.W. Hong and Y.L. Chen, *Impact of using an educational robot-based learning system on students motivation in elementary education*, IEEE Transactions on learning technologies, 7(4), 2014.

[3] M. Alemi, A. Meghdari and M. Ghazisaedy, *Employing humanoid robots for teaching english language in Iranian junior high-schools. International Journal of Humanoid Robotics*, 11(03), 2014.

[4] A. Kapoor, and R.W. Picard, *Multimodal affect recognition in learning environments*, In Proceedings of the 13th annual ACM international conference on Multimedia. ACM, 2005.

[5] D. Leyzberg, S. Spaulding and B. Scassellati, *Personalizing robot tutors to individuals' learning differences*, In Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction. ACM, 2014.

[6] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati, *The physical presence of a robot tutor increases cognitive learning gains*, 2012.

[7] J. Kennedy, P. Baxter, T. Belpaeme, *The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning*, In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Mar 2 . ACM, 2015.

[8] R.S. Baker, S.K. D'Mello, M.M.T. Rodrigo and A.C. Graesser, *Better to be frustrated than bored: The incidence, persistence, and impact of learners cognitiveaffective states during interactions with three different computer-based learning environments*, International Journal of Human-Computer Studies, 68(4), 2010.

[9] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati, *The physical presence of a robot tutor increases cognitive learning gains*, 2012.

[10] J. Kennedy, P. Baxter, T. Belpaeme, *The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning*, In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Mar 2 . ACM, 2015.

[11] R.S. Baker, S.K. D'Mello, M.M.T. Rodrigo and A.C. Graesser, *Better to be frustrated than bored: The incidence, persistence, and impact of learners cognitiveaffective states during interactions with three different computer-based learning environments*, International Journal of Human-Computer Studies, 68(4), 2010.

[12] M.E. Hoque, D.J. McDuff and R.W. Picard, *Exploring temporal patterns in classifying frustrated and delighted smiles*, IEEE Transactions on Affective Computing, 3(3), 2012.

[13] Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., and Lester, J.C. Automatically recognizing facial indicators of frustration: A learning-centric analysis. (2013).

[1] Each emotion instance in Cohn-Kanade is represented by a series of images, starting from neutral, peaking to the most representative state of the emotion and then returning back to neutral. Peaked images are the ones, roughly in the centre of each series, in which the expression of the emotion has peaked to the most representative state for each test subject.

[14] Littlewort, G., Whitehill, J., Wu, T., et al. The computer expression recognition toolbox (CERT). 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011), (2011), 298305.

[15] Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., and Movellan, J.R. The faces of engagement: Automatic recognition of student engagement from facial expressions. IEEE Transactions on Affective Computing 5, 1 (2014), 8698.

[16] Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2015. Automatic Detection of Learning-Centered Affective States in the Wild. In Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15). ACM, New York, NY, USA, 379-388.

[17] P. Lucey, J. Cohn, T. Kanade, J. Saragih and Z. Ambadar, "A complete expression dataset for action unit and emotion-specified expression," in Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis, San Francisco, 2010.

[18] Facial Action Coding System: A Technique for the Measurement of Facial Movement. Ekman, P. and Friesen, W. Palo Alto : s.n., 1978. Consulting Psychologists Press.