Proceedings of the 2017 Design of Medical Devices Conference
DMD2017
April 10-13, 2017, Minneapolis, Minnesota, USA
DMD2017-3332

Multimodal Affect Recognition for Assistive Human-Robot Interactions

Alexander Hong, Yuma Tsuboi, Goldie Nejat, Beno Benhabib

Department of Mechanical & Industrial Engineering, University of Toronto, Toronto, Canada

1 Introduction

Socially assistive robots can provide cognitive assistance with activities of daily living, and promote social interactions to those suffering from cognitive impairments and/or social disorders. They can be used as aids for a number of different populations including those living with dementia or autism spectrum disorder, and for stroke patients during post-stroke rehabilitation [1]. Our research focuses on developing socially assistive intelligent robots capable of partaking in natural human-robot interactions (HRI). In particular, we have been working on the emotional aspects of the interactions to provide engaging settings, which in turn lead to better acceptance by the intended users. Herein, we present a novel multimodal affect recognition system for the robot Luke, Fig. 1(a), to engage in emotional assistive interactions.

Current multimodal affect recognition systems mainly focus on inputs from facial expressions and vocal intonation [2], [3]. Body language has also been used to determine human affect during social interactions, but has yet to be explored in the development of multimodal recognition systems. Body language has been strongly correlated to vocal intonation [4]. The combined modalities provide emotional information due to the temporal development underlying the neural interaction in audiovisual perception [5].

In this paper, we present a novel multimodal recognition system that uniquely combines inputs from both body language and vocal intonation in order to autonomously determine user affect during assistive HRI.

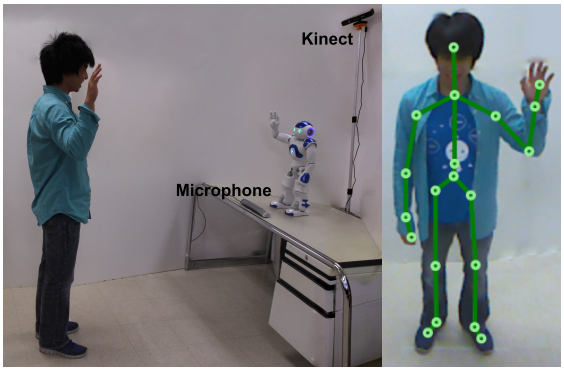

Fig. 1. (*a*) Emotional Interactions with the Luke Robot and a User, and (*b*) Body Pose Tracking using the Kinect Sensor.

2 Body Language and Vocal Intonation

Both body language and vocal intonation convey information about a person's affect during social interaction [4]. Vocal intonation is strongly influenced by body posture and movement due to changes in the vocal tract [5]. Affect from body language is found to be biased towards the affective information simultaneously presented through vocal intonation, implying that body language and tone of voice are perceptually combined into an audiovisual percept [6].

This work focuses on investigating the autonomous recognition of affect in real-time by uniquely combining a user's body language and vocal intonation during HRI. We use the 2D valence-arousal scale to determine affect due to its encompassing of all possible affective states and their variations [7]. Valence and arousal are used to measure a user's pleasure level and level of excitation, respectively [8].

3 Multimodal Affect Recognition System

Our proposed automated multimodal affect recognition system consists of three main sub-systems, Fig. 2: (1) body language detection and classification, (2) vocal intonation detection and classification, and (3) multimodal affect identification. We utilize a decision-level fusion approach to estimate user's affect from both modes.

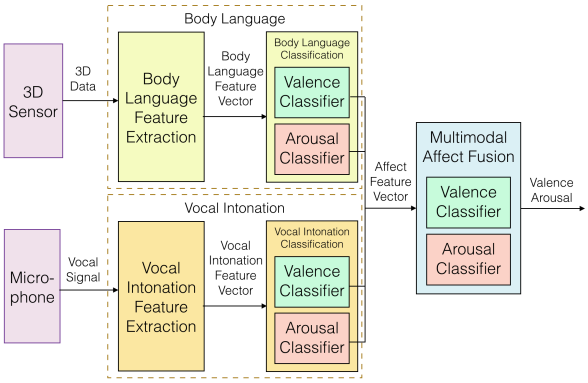

Fig. 2. Multimodal Affect Recognition System.

3.1 Body Language Features

The body language features, utilized herein, were adapted from our previous work, and have been shown to be directly correlated to a person's affect [9]. They consist of bowing/stretching of the trunk, opening/closing of the arms, vertical head position and motion of the body, forward/backwards head position and motion of the body, expansiveness of the body, and speed of the body. The features are extracted from 20 position coordinates (head, shoulder center, spine, hip center, both wrists, elbows, shoulders, hips, knees, ankles, feet, and left and right hands). The features are tracked in real-time using 3D information provided by the Kinect 3D sensor, Fig. 1(b). A feature vector is generated at a sampling rate of 30 fps for body language classification. A classifier is, then, used to obtain the body language valence and arousal levels of the user.

3.2 Vocal Intonation Features

We utilize 51 vocal intonation features to classify valence and arousal of a vocal utterance. These features were selected based on their influence on vocal intonation, and their characterization of the physiological changes in respiration – details of each feature are described in Table 1.

In order to extract the vocal intonation features during

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Research Chairs (CRC) Program.

HRI, the Voice Tracker II noise-cancelling microphone array (Fig. 1a) was used to record the user's utterance in real-time in two second segments. Once the feature vector is obtained, a classifier is used to identify valence and arousal levels.

Table 1. Vocal Intonation Feature Definition.

Vocal Feature	Definition [10]
Zero Crossing Rate (ZCR)	ZCR is the rate of sign-changes along a signal. It indicates the proportions of the speech signal that are voiced or unvoiced.
Spectral Centroid Spread (SCS)	SCS is the weighted mean of the frequencies in the signal with frequency magnitudes as weights. SCS can characterize the physiological changes in articulation.
Spectral Flux	Spectral flux is the rate of change of the power spectral density of a signal.
18 Mel-frequency Cepstral Coefficients (MFCC)	MFCC represents the Mel-frequency cepstrum, a short-term power spectrum of a sound based on a linear cosine transform of a log power spectrum on a nonlinear Mel-scale of frequency. MFCC characterizes the vocal tract configuration.
Fundamental Frequency (F_0)	Fundamental frequency is the lowest frequency of a periodic waveform. Fundamental frequency, or pitch, is directly related to emotions in speech.
Chroma Vector of 12 Values	Chroma vector is a pitch distribution feature that describes tonality, measuring the energy of each of the 12 pitch classes of the equal-tempered scale within an analysis frame.
Energy	Energy is the magnitude of the signal, or the area under the signal curve, and it can indicate the proportion of voice segments that are voiced and unvoiced.
Spectral Entropy	Spectral entropy is the measure of the spectral distribution to detect voiced and unvoiced components of speech. It quantifies the degree of randomness of the spectral probability density.
Spectral Roll Off	Spectral roll off is the N^{th} percentile of the power spectral distribution of the audio signal, where N is usually chosen to be 85-95%.
14 Linear Prediction Coding (LPC) Coefficients	LPC coefficients approximate the current speech signal as a linear combination of previous speech signals. They characterize the speaker's vocal tract shape.

3.3 Multimodal Affect Fusion

The classified valence, v_b and v_v, and arousal, a_b and a_v, values from body language and vocal intonation are combined to form an affect feature vector for decision-level fusion, i.e., $\mathbf{c}_m = \begin{bmatrix} v_m \\ a_m \end{bmatrix}$. Both the classified body language and vocal intonation affective values have a one-to-one correspondence with each other, as they are acquired during the same 2 s interaction time interval.

4 Experiments

In order to validate the proposed multimodal affect recognition system, the body language detection and classification, vocal intonation detection and classification, and multimodal affect fusion sub-systems were each trained utilizing a database we created with 446 corresponding body language and vocal intonation instances obtained from actors displaying both body language and vocal intonation.

A 10-fold cross-validation method was used to validate the use of different classifiers. These classifiers encompass a variety of learning techniques, including probabilistic learning, decision trees, lazy learning algorithms, neural networks, and non-linear models. The classifiers used and their results are presented in Table 2. Random forest decision trees provided the highest body language classification rates of 93.6%, and 95.2% for valence and arousal, respectively. Multi-layer perceptron neural networks achieved the highest vocal intonation classification rates of 79.4%, and 84.4% for valence and arousal, respectively. A Bayesian network was determined to provide the highest multimodal classification rates of 96.0%, and 98.2% for valence and arousal, respectively, Table 2. SVM also achieved a recognition rate of 96.0% for multimodal valence.

Table 2. 10-fold Cross-validation Results for Affect Classification.

Classifier	Classification Rate (%)					
	Body Language		Vocal Intonation		Multimodal	
	Valence	Arousal	Valence	Arousal	Valence	Arousal
Bayesian Network	86.2	89.7	58.0	72.0	**96.0**	**98.2**
Naïve Bayes	81.0	89.0	54.5	70.9	94.4	97.1
Logistic Regression	84.9	91.8	71.4	75.7	94.4	96.8
Random Forest	**93.6**	**95.2**	70.3	78.6	95.5	98.0
k-nearest neighbors	92.9	94.5	71.5	76.1	95.2	97.8
Multi-layer perceptron	92.9	92.4	**79.4**	**84.4**	95.1	97.5
Support Vector Machines	84.0	88.1	75.8	80.9	**96.0**	98.0

5 Interpretation

The objective of our research is to develop an intelligent socially assistive robot to assist people with activities of daily living and provide social interventions. In this paper, we presented an automated multimodal affect recognition and classification system for determining a user's affect levels from body language and vocal intonation information during HRI. Our results show that by using a multimodal decision-level fusion, we can obtain higher recognition rates than when using the individual modalities on their own. Future work will consist of investigating the addition of other modes (e.g., facial expressions and physiological signals) into our multimodal affect recognition system to investigate the impact of different affective modalities during HRI with an assistive robot. The robot will utilize the user's affect in order to determine its own appropriate assistive behaviors.

References

[1] Tapus, A., Matarić, M.J. and Scassellati, B., 2007, "Socially assistive robotics," IEEE Robot. Autom. Mag, 14(1), pp. 35-42.

[2] McColl, D., Hong, A., Hatakeyama, N., Nejat, G., and Benhabib, B., 2016, "A survey of autonomous human affect detection methods for social robots engaged in natural HRI," J. Intell. Robot. Syst., 82(1), pp. 101-133.

[3] McColl, D., and Nejat, G., 2014, "Recognizing Emotional Body Language Displayed by a Human-like Social Robot," Int J of Soc Robotics, 6(2), pp. 261–280.

[4] Van den Stock, J., Righart, R., and de Gelder, B., 2007, "Body expressions influence recognition of emotions in the face and voice," Emotion, 7(3), pp. 487-494.

[5] Jessen, S., Obleser, J., and Kotz, S., 2012, "How Bodies and Voices Interact in Early Emotion Perception," PLoS ONE, 7(4), p. e36070.

[6] de Gelder, B., de Borst, A., and Watson, R., 2014, "The perception of emotion in body expressions," Wiley Interdiscip Rev Cogn Sci, 6(2), pp. 149-158.

[7] Barrett, L., 1998, "Discrete Emotions or Dimensions? The Role of Valence Focus and Arousal Focus," Cogn. Emot., 12(4), pp. 579-599.

[8] Russell, J., Weiss, A., and Mendelsohn, G., 1989, "Affect Grid: A single-item scale of pleasure and arousal," J. Pers. and Soc. Psychol., 57(3), pp. 493-502.

[9] McColl, D., and Nejat, G., 2014, "Determining the Affective Body Language of Older Adults during Socially Assistive HRI," Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., Chicago, IL, pp. 2633-2638.

[10] Schuller, B., 2013, "Audio Features," Intelligent Audio Analysis, Signals and Communication Technology, pp 41-97.