Interactive Robot Learning for Multimodal Emotion Recognition

Chuang Yu, Adriana Tapus

Interactive Robot Learning for Multimodal Emotion Recognition *

Chuang Yu[1] and Adriana Tapus[1]

Autonomous System and Robotics Lab, U2IS, ENSTA Paris
Institut Polytechnique de Paris
828 boulevard des Marchaux, 91120 Palaiseau, France
E-mail: {chuang.yu; adriana.tapus}@ensta-paristech.fr

Abstract. Interaction plays a critical role in skills learning for natural communication. In human-robot interaction (HRI), robots can get feedback during the interaction to improve their social abilities. In this context, we propose an interactive robot learning framework using multimodal data from thermal facial images and human gait data for online emotion recognition. We also propose a new decision-level fusion method for the multimodal classification using Random Forest (RF) model. Our hybrid online emotion recognition model focuses on the detection of four human emotions (i.e., neutral, happiness, angry, and sadness). After conducting offline training and testing with the hybrid model, the accuracy of the online emotion recognition system is more than 10% lower than the offline one. In order to improve our system, the human verbal feedback is injected into the robot interactive learning. With the new online emotion recognition system, a 12.5% accuracy increase compared with the online system without interactive robot learning is obtained.

Keywords: Interactive robot learning · Multimodal emotion recognition · Human-robot interaction.

1 Introduction

In the past decade, emotion detection during social interaction attracted more and more attention with the rapid advances in the field of robotics. Research studies have found human emotion perception as a fundamental component of communication that plays a significant role in successful human-human interaction [8]. Emotion recognition during human-robot interaction can help robots to understand user's state and exhibit a natural social interaction. However, human emotion recognition is challenging. Many researchers use multimodal information to address emotion recognition [13]. In our paper, we also use a multimodal system combining the information from thermal face images and gait information during human-robot interaction. Even though many researchers study emotion recognition models for robots, it is very expensive and time-consuming to label and annotate large databases by hand. Interactive Robot Learning (IRL) can

* Supported by ENSTA Paris.

address this problem. During human-robot interaction, robots can get verbal feedback from humans to label or relabel the data extracted from the interaction [7]. IRL is very useful in the long-life learning situation where there is no large-scale data for emotion recognition. Robots can record the emotion-related features and obtain its label from the interaction with humans. Currently, most of the researches focus on offline emotion analysis. However, the online recognition capability is more challenging. The robot interactive learning can also improve flexibility of emotion recognition model in human-robot interaction.

In the last decades, numerous studies attempted the multimodal method with the visual, verbal, and physiological signals and the natural language in order to get a better ability of emotion recognition. Caridakis et al. [2] made use of the multimodal Bayesian classifier with features extracted from face images, gestures, body action, and speech information to classify 8 emotions namely anger, despair, interest, pleasure, sadness, irritation, joy, and pride. In addition, they employed the fusion methods at feature-level and decision-level. Regarding IRL, authors in [6] described what was IRL with mixed-initiative and how the memory-based human-robot interaction strategies worked in the learning environment. Lutkebohle et. al [9] developed an IRL system with the human speech feedback in the dialog loops to help a curious robot learn skills in the grasping task. The human feedback from the human-robot interaction was proved significantly useful during robot learning.

Moreover, to the best of our knowledge, no studies focus on the fusion of gait and thermal facial features together to detect human emotion in social robotics context. A very common scenario is when the human walks towards the social robot and stops in front of it to interact with it. In our work, we developed a multimodal emotion recognition method by using thermal facial images during human-robot interaction and gait data during walking towards the robot to recognize four emotional states (i.e., neutral, happy, angry, and sad). The offline emotion recognition is widely developed by researchers. However, the online testing of emotion recognition model is challenging in real-time HRI context. In this paper, we developed a new method based on Random Forest (RF) model and confusion matrices of two individual Random Forest models by using the data from the face thermal images and gait. In addition, the IRL method is used with the verbal human feedback in the learning loops in order to improve the performance of real-time emotion recognition. The experiment results shows the effectiveness of IRL in multimodal emotion recognition.

The rest of the paper is structured as follows: Section 2 describes the methodology; Section 3 shows the experimental setup. The experimental results are summarized in Section 4. The conclusions and discussions are part of Section 5.

2 Methodology

2.1 Multimodal database

In our study, we collected the data (i.e., the thermal facial images and gait data) from multiple modalities. The database was built up with the data from

the human-robot interaction experiments. Here, we use it for the training of the offline emotion recognition model. Many multimodal databases for affective computing exist in the literature. However, to the best of our knowledge, there are no open multimodal databases with face thermal images and gait data. In our past research, we conducted 300 human-robot interaction experiments with 15 participants to build up our database. During the experiments, each participant run 20 experiments (i.e., 5 times each emotion). The gait data was extracted from the RGB-D camera during human walking towards the robot and included 3D skeleton data with the positions of 25 joints and related timestamps of each frame. The trajectories of 3D skeletons are shown in Fig. 1. The thermal images with the human upper body were recorded from the thermal camera Optris Pi during human-robot interaction for 10 seconds. Then, we selected the suitable data where the participants' heads did not move much. The time duration of all data in the thermal database is from 5 seconds to 10 seconds. The examples of the thermal database are shown in Fig. 2.

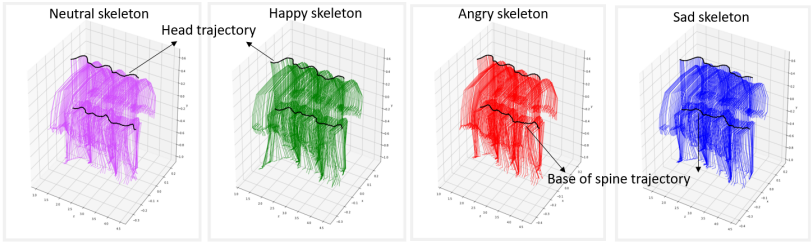

Fig. 1. The trajectories of skeleton

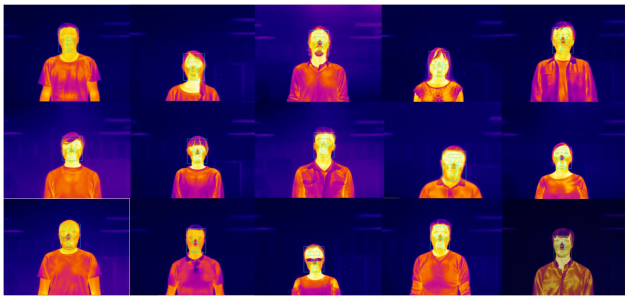

Fig. 2. Examples of thermal images from the database

2.2 Features extraction

From the RGB-D camera, we extracted the 3D positions of 25 joints. The index of the 25 joints in the human body is shown in Fig. 3. In this paper, 4 joint angles and 4 joint angular velocities are calculated as features including the left knee joint angle (LKJA) and its velocity (LKJAV), the right knee joint angle (RKJA) and its velocity (RKJAV), the left hip joint angle (LHJA) and its velocity (LHJAV), the right hip joint angle (RHJA) and its velocity (RHJAV). The definition of the joint angles is shown in Fig. 2.2. LHJA is calculated from 3D positions of joints 1, 0, 12, and 13. LKJA is calculated from 3D positions of joints 12, 13, and 14. RHJA is calculated from 3D positions of joints 1, 0, 16, and 17. RKJA is calculated from 3D positions of joints 16, 17, and 18. LHJAV, LKJAV, RHJAV, and RKJAV are just calculated through a subtraction operation of the current angle and the previous angle and the following division operation on the sampling time gap. The more detailed description of the calculation of the 8 parameters was presented in one of our past work [3]. In this study, we use Power Spectral Density (PSD) as the feature of the gait data, which describes the energy of the specific frequency or frequency range. Then, Welch method is applied to calculate PSD features [11].

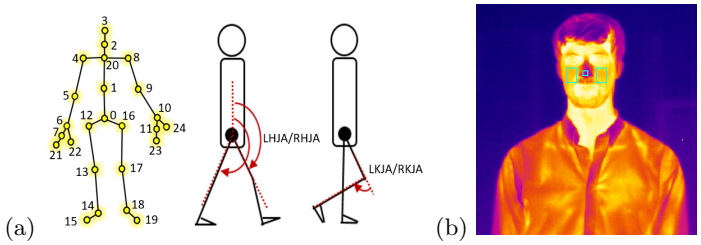

Fig. 3. Data: (a) Joint index and joint angle definition and (b) 3 facial ROIs

Many researchers have found that human emotions are correlated to the facial areas (e.g., left cheek, right cheek, and nose [14] [10]). In this paper, we use these 3 facial Region of Interest (ROI) for human emotion detection [1] (see Fig. 3 (b)). The average value and variance of these 3 facial regions are used as thermal facial features.

2.3 Fusion of multimodal classifiers

Previously, we have tested many machine learning models for offline emotion recognition including Hidden Markov Model (HMM) with thermal data, HMM model with gait data, and Random Forest (RF) model with thermal data, RF model with gait data, Controversial Neural Network (CNN) model with gait data, Support Vector Machine (SVM) model with thermal data, SVM model with gait data. In addition, CNN model with gait feature has only an offline

testing accuracy of 55%. The offline accuracy of HMM, SVM and RF with gait data are 65%, 65% and 70%, respectively. SVM and RF have the accuracy of 55% and 60%, respectively. SVM and RF with both features obtain 70% and 80%, respectively. Hence, we found that RF had the best emotion recognition performance with the highest accuracy with respect to the other models. So, this paper applies RF as the basic machine learning model for emotion recognition.

The individual emotion classifier only with gait data and the one only with thermal face data have different recognition abilities for the 4 different emotions. During online testing, we found that the two individual models show distinct emotion recognition performance with different accuracy in each emotion situation. Hence, the decision-level fusion of the two individual models is necessary to make a better recognition accuracy. In this paper, we developed a new decision-level fusion method with two RF classifiers for online emotion recognition and the framework is as shown in Fig. 4. The integration method is based on the modified confusion matrix and the probability vector of each emotion class.

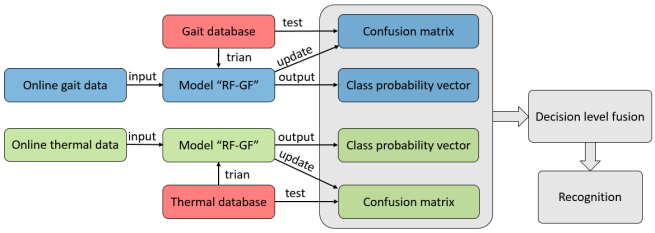

Fig. 4. The fusion framework of multimodal classifiers

We use the confusion matrix information to build up the decision-level hybrid model for emotion recognition. In the confusion matrix all the elements represent probabilities. The row of the confusion matrix represents the instances in a predicted class while the column represents the instances in the real class. Then, the elements of each column in the confusion matrix is divided by the total amount of instances in each column, respectively to get the modified confusion matrix. In the modified one, each column shows the probability of each real class in the predicted class situation. An example of our modified confusion matrix for the thermal emotion recognition model is shown in Fig. 5. For example, in column two corresponding to the predicted class angry in Fig. 5, TC_{02}, TC_{12}, TC_{22}, and TC_{32} represent the probabilities of neutral, happy, angry, and sad emotions, respectively when the predicted class is the angry emotion.

During every online testing, the class probabilities of the thermal model make up the vector $TPM_{1\times4}$, as shown in Fig. 6. For example, the thermal emotion recognition model gets the prediction result happy. From the above matrix, the predicted happy is neutral with probability TC_{01}, happy with probability TC_{11}, angry with probability TC_{21}, and sad with probability TC_{31}. Therefore, in the

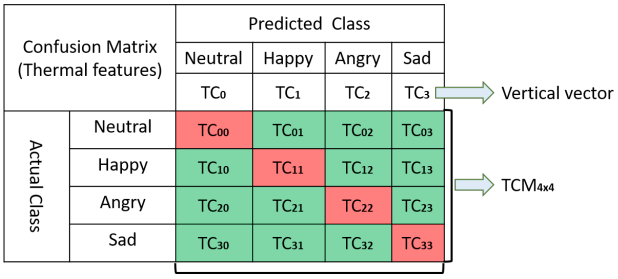

Fig. 5. Modified confusion matrix of thermal emotion recognition model

happy prediction situation, the probability vector of the 4 emotions is equal to $TP_1 \times TC_1$. Similarly, we can get the probability vectors of the 4 emotions in other 3 prediction situations. The process is as shown in Fig. 6.

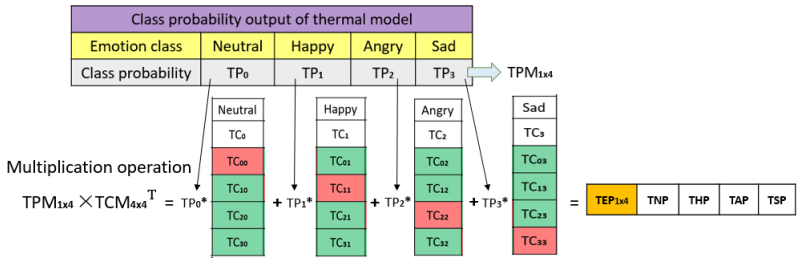

Fig. 6. Class probability calculation with modified confusion matrix

The calculation of $TEP_{1\times4}$ for the thermal model, of $GEP_{1\times4}$ for the gait model, of $FUEP_{1\times4}$ for the two new vectors, and of FRR for the final recognition result is as indicated in Eqs. 1, 2, 3, and 4, respectively.

$$TEP_{1\times4} = TPM_{1\times4} \times TCM_{4\times4} \tag{1}$$

$$GEP_{1\times4} = GPM_{1\times4} \times GCM_{4\times4} \tag{2}$$

$$FUEP_{1\times4} = TEP_{1\times4} + GEP_{1\times4} \tag{3}$$

$$FRR = argmax(FUEP_{1\times4}) \tag{4}$$

2.4 Robot interactive learning

We apply an interactive learning method during the interaction in order to boost the emotion recognition performance of the robot. An overview of IRL system architecture is illustrated in Fig. 7. During the IRL experiments, if the predicted emotion does not match the real one, the gait and the thermal facial features

are restored to retrain the emotion recognition models. The updated thermal and gait models will test the saved features again to get the new predicted results, respectively. The new predicted results will help update the two confusion matrices. If the predicted emotion is equal to the real one, we only update the two confusion matrices and the models are not retrained.

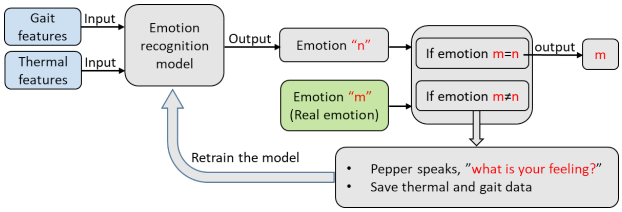

Fig. 7. Overview of the IRL architecture

3 Experimental design

In our work, we use Kinect camera to detect the human gait information and the Optris thermal camera to obtain the thermal facial images for emotion recognition. In our experiments, we used Pepper robot. The experimental setup is as shown in Fig. 8. Our experiment is composed of three parts: the online testing,

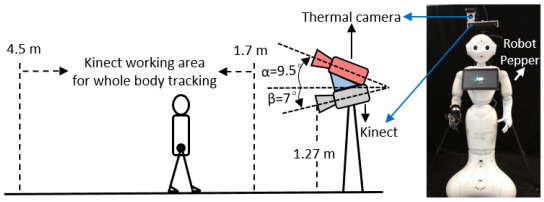

Fig. 8. Experimental setup

the IRL, and the online testing after IRL. 8 participants took part 8 times in each part. Hence, 64 experiments in every condition and 192 experiments in total were conducted for one month. The participants are students with different major backgrounds, with ages from 20 to 32, half women and half men. All the experiments were conducted in our robotics lab, with all the windows closed and the air conditioning turned on in order to keep the indoor temperature ranging from 20 to 28 degrees Celsius.

For the emotion elicitation methods, static images and films clips are considered extensively as stimuli to elicit different emotions in the laboratory and

films are one of the most effective ways to elicit emotions [4]. In our experiments, for emotion elicitation, we used open film clips database-FilmStim [12], in which there are English and French videos that can elicit emotional states including amusement, anger, sadness, tenderness, fear, disgust, and neutral state. In our research, we only focus on 4 emotions: neutral, amusement, anger, and sadness. Hence, the film clips of 4 emotions in the database are used in our research, and more precisely: 6 clips for neutral emotion (3 English, 3 French), 13 clips for happy emotion (3 English, 10 French), 16 clips for angry emotion (9 English, 7 French), and 17 clips for sad emotion (7 English, 10 French). The film clips for each emotion were randomly chosen for emotion elicitation. After emotion elicitation experiments, we applied Pick-A-Mood (PAM) [5] as a tool of emotional state measure.

Each experiment is composed of three parts: (1) the emotion elicitation part, (2) the walking part for emotional gait data, and (3) the standing part for thermal facial images. In the emotion elicitation part, the participant randomly selects and watches one or more film clips for one or more times for a specific emotion. Then, the participant walks towards the robot with that specific emotion and stops before the robot to start the interaction. The details of IRL part are described in the following steps. For the other two parts, before and after IRL, there is no step (6). The steps are as described below: (1) The participant completes EPQ to determine his/her personality traits (one time); (2) The participant randomly selects and watches the film clips from FilmStim for a specific emotion; (3) The participant selects the emotional state from PAM for mood measurement; (4) The robot instructs the participant to walk towards it with the specific emotional gait; (5) The participant walks towards the robot and stops before it for 5 seconds; (6) The robot asks, how do you feel? The robot records the emotion result and gait/thermal data into a database for retraining the emotion recognition model; (7) Repeat steps from 2 to 6 to complete eight or more experiments.

4 Experimental results

Before online testing for emotion recognition, the offline RF models were tested in order to get modified confusion matrix, which is used for fusion of our emotion recognition models. There were 80 testing conducted in this part. The modified confusion matrix shows the probability distribution relation of the four emotions as shown in Table 1 and Table 2, respectively. In the modified confusion matrix, each column sums up to 100% (if without rounding) instead of each row in the original confusion matrix. And, the primary diagonal elements in the confusion matrix represent accuracy of each emotion while modified confusion matrix's diagonal elements just indicate a conditional probability instead of accuracy of each emotion. In Table 2, the ratio with 91.67% is significantly high when the real class and the predicted class both are sad one. That ratio in the modified confusion matrix only represents the probability of "sad" when predicted class is "sad", instead of the accuracy of "sad" (only 55%) in the confusion matrix.

Table 1. Modified confusion matrix of offline testing of gait model

Gait model (Offline testing)		Predicted class			
		Neutral	Happy	Angry	Sad
Real class	Neutral	75%	16%	0%	6.67%
	Happy	0%	52%	30%	6.67%
	Angry	10%	20%	65%	0%
	Sad	15%	12%	5%	86.67%

Table 2. Modified confusion matrix of offline testing of thermal models

Thermal model (Offline testing)		Predicted class			
		Neutral	Happy	Angry	Sad
Real class	Neutral	52%	25%	5.26%	0%
	Happy	12%	50%	21.05%	8.33%
	Angry	16%	20.83%	57.89%	0%
	Sad	20%	4.17%	15.79%	91.67%

For the integration of the model with gait features and the one with thermal facial features, we applied a decision-level method with the modified confusion matrix. Before IRL experiments, we conducted the online testing experiments with the decision-level model. The paper compared the recognition performance of the single models with gait features or thermal facial features and the multimodal model with both, which are with the accuracy of 54.6875% (gait model), 59.375% (thermal model), and 65.625% (fusion model), respectively. The results indicate a higher accuracy for the hybrid model than for any single models.

During IRL, when hybrid model recognition result does not match the real one, the data are saved for model retraining. The confusion matrices are updated with the testing results on the two individual classification models, which are retrained. After IRL, the online testing was conducted again with the updated hybrid model. At last, an online testing accuracy of 78.125% is obtained after IRL, which is an increase of more than 10% than the one before interactive learning (i.e., 65.625%), which demonstrates that the IRL method is useful for emotion recognition with gait and thermal facial data.

5 Conclusion and Discussion

In our paper, we developed a multimodal emotion recognition model with gait and thermal facial data, which is based on RF model and the modified confusion matrices of two individual models. Through comparison between the individual RF models and the decision-level hybrid model, we found that our integration method is useful to classify the emotion during human-robot interaction. In addition, we have conducted 192 experiments including three parts, namely online testing experiments before IRL, IRL experiments, and online testing after IRL. In these experiments, we compared the emotion recognition performance before and after IRL to find out that the interactive robot learning is useful, with an

increase of more than 10% of the accuracy of multimodal emotion recognition with gait and thermal data. Our IRL architecture could potentially be used in robot long life learning in human-robot interaction scenarios.

References

1. Agrigoroaie, R., Tapus, A.: Physiological parameters variation based on the sensory stimuli used by a robot in a news reading task. In: 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). pp. 618–625. IEEE (2018)
2. Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., Karpouzis, K.: Multimodal emotion recognition from expressive faces, body gestures and speech. In: IFIP International Conference on Artificial Intelligence Applications and Innovations. pp. 375–388. Springer (2007)
3. Chuang, Y., Adriana, T.: Multimodal emotion recognition with thermal and rgb-d cameras for human-robot interaction. In: 2019 19th International Conference on Advanced Robotics (ICAR). IEEE (2019 Under review)
4. Deng, Y., Yang, M., Zhou, R.: A new standardized emotional film database for asian culture. Frontiers in psychology **8**, 1941 (2017)
5. Desmet, P.M., Vastenburg, M., Romero, N.: Mood measurement with pick-a-mood: review of current methods and design of a pictorial self-report scale. Journal of Design Research **14**(3), 241–279 (2016)
6. Hanheide, M., Sagerer, G.: Active memory-based interaction strategies for learning-enabling behaviors. In: RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication. pp. 101–106. IEEE (2008)
7. Katagami, D., Yamada, S.: Interactive classifier system for real robot learning. In: Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication. IEEE RO-MAN 2000 (Cat. No. 00TH8499). pp. 258–263. IEEE (2000)
8. Keltner, D., Haidt, J.: Social functions of emotions at four levels of analysis. Cognition & Emotion **13**(5), 505–521 (1999)
9. Lutkebohle, I., Peltason, J., Schillingmann, L., Wrede, B., Wachsmuth, S., Elbrechter, C., Haschke, R.: The curious robot-structuring interactive robot learning. In: 2009 IEEE International Conference on Robotics and Automation. pp. 4156–4162. IEEE (2009)
10. Nakanishi, R., Imai-Matsumura, K.: Facial skin temperature decreases in infants with joyful expression. Infant Behavior and Development **31**(1), 137–144 (2008)
11. Rahi, P., Mehra, R.: Analysis of power spectrum estimation using welch method for various window techniques. International Journal of Emerging Technologies and Engineering **2**(6), 106–109 (2014)
12. Schaefer, A., Nils, F., Sanchez, X., Philippot, P.: Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. Cognition and Emotion **24**(7), 1153–1172 (2010)
13. Sebe, N., Cohen, I., Gevers, T., Huang, T., et al.: Multimodal approaches for emotion recognition: A survey (2005)
14. Sugimoto, Y., Yoshitomi, Y., Tomita, S.: A method for detecting transitions of emotional states using a thermal facial image based on a synthesis of facial expressions. Robotics and Autonomous Systems **31**(3), 147–160 (2000)