UK Research and Innovation

UKRI TAS Node on Trust

Affective XAI for Trustworthy Human-Robot Interaction

Chuang Yu, Hongbo Zhu, Angelo Cangelosi

UKRI Node on Trust

Cognitive Robotics Lab, CS Department, University of Manchester

Abstract

With the improvement of computing power and the availability of large datasets, deep learning models can achieve excellent performance in facial expression recognition tasks. As these deep neural networks have very complex nonlinear structures, it isn't easy to understand the basis for the model's prediction when it makes a prediction. Specifically, we do not know which facial features contribute to the classification. The affective computing models with more explainable and transparent feedback for human interactors are essential for a trustworthy human-robot interaction (HRI). This paper explores facial emotion recognition with an explainable AI (XAI) model in a human-robot interaction scene. The experiment with the Pepper robot shows promising results regarding the visual and verbal feedback with the affective explainability from XAI. The affective XAI model can explain the facial expression prediction made by the emotion recognition model at the pixel level and why the emotion is recognized at the face action unit level, which facilitates HRI to be more natural and reliable.

Pipeline of our framework

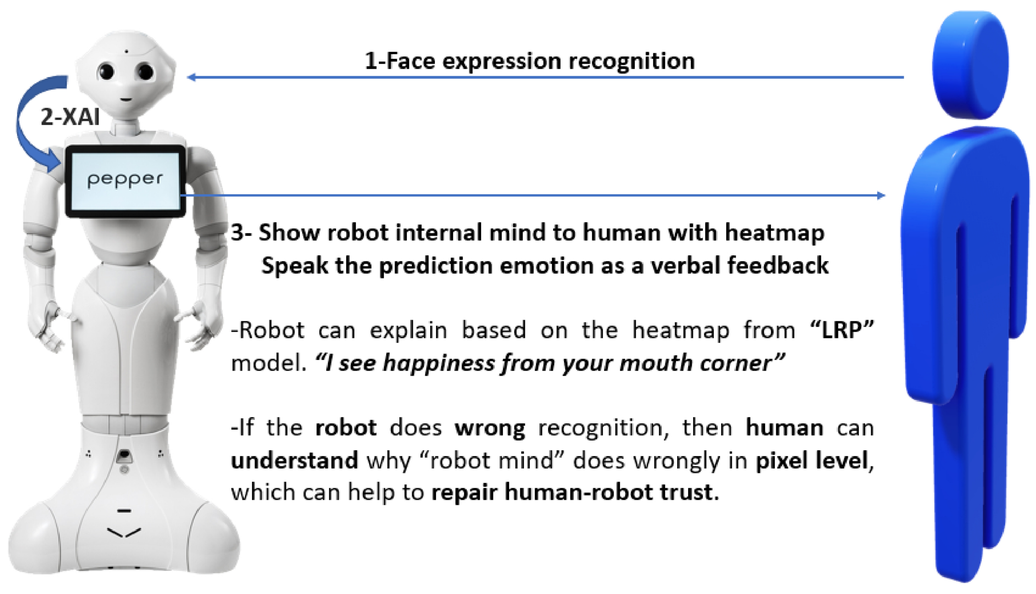

Fig. 1. Pipeline

1. Firstly, the Pepper robot predicts the facial emotional states of the interactor during HRI.
2. Then, the robot extracts the explainable presentation with the help of the XAI model.
3. Pepper verbalises the predicted emotion as the linguistic feedback and shows the heatmap extracted from the LRP model as explainable visual feedback.
4. In addition, the robot can give more detailed emotion recognition feedback to increase the interaction transparency. This explainable visual feedback can help the human interactor understand the internal process of the robot's emotion recognition model, which in turn facilitates human-robot trust.

Human Emotion recognition Model

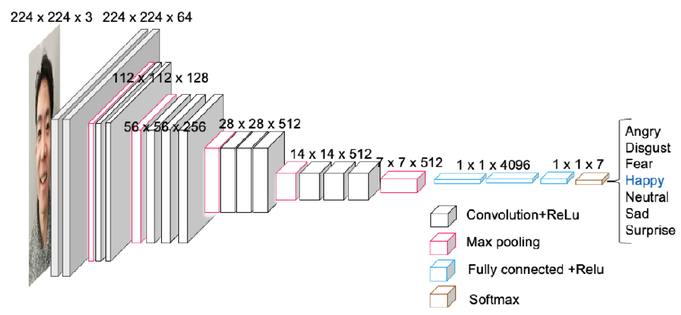

Fig. 2. Emotion recognition model based on VGG16

Fig. 3. KDEF Dataset
(Karolinska Directed Emotional Faces)

XAI model

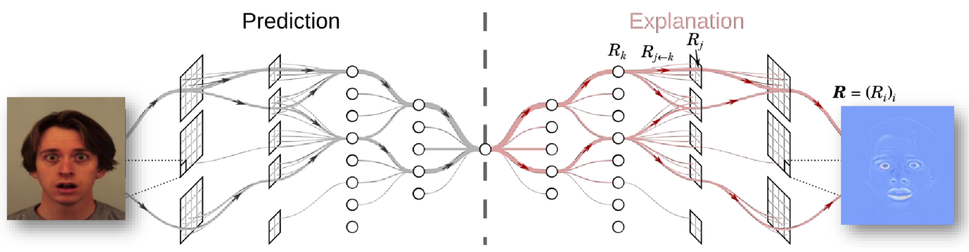

Fig. 4. XAI model based on Wise Relevance Propagation (LRP)
(Red arrows indicate the direction of back propagation flow)

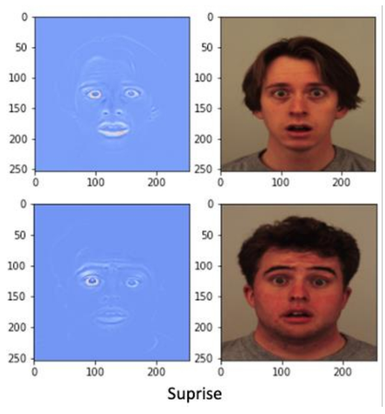

Fig. 5. XAI results with heatmap

Training and Testing and Explainable Heatmap

- Based on pre-trained VGG16 model, our face emotion recognition model is further trained on an Nvidia RTX 2080Ti graphic card. (batch size=32; Adam for optimization; learning rate = 0.00001.
- After 250 epochs training, testing accuracy of 91.4% on KDEF dataset.
- Pixel-level explanation with heatmap through XAI as shown in Fig. 5.

XAI in Human-Robot Interaction

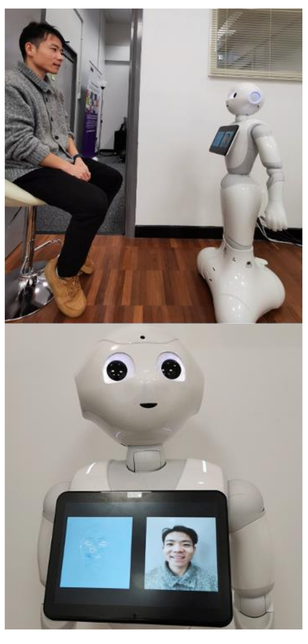

Fig. 6. HRI scene with XAI model

- Verbal feedback: Peper can recognise user facial expression and verbalise the human face emotion prediction as verbal feedback in HRI. For example, robot will speak *"you looks like happy!"*. The speech voice is synthesized through the Text-To-Speech (TTS) tool of the Naoqi SDK of the Pepper robot.
- Visual feedback: Based on XAI model, Pepper robot can extract the heatmap images as the pixel-level explanation shown on Pepper screen as explainable visual feedback.
- Trustworthy HRI: Through verbal prediction feedback and explainable visual feedback, it is promising to build up a trustworthy HRI.

HERIOT WATT UNIVERSITY | Imperial College London | MANCHESTER 1824 The University of Manchester | UKRI Trustworthy Autonomous Systems Hub