北京郵電大學

Queen Mary
University of London

Undergraduate Project Report
2018/19

# Facial emotion recognition applied to a humanoid robot teacher

**Date: 6 May, 2019**

# Table of Contents

# Abstract

Humanoid robot has emerged as a useful tool to improve students' engagement and sustain learning. However, in most robotics education scenarios, the state of students is overlooked, so that effective learning may not be guaranteed. To address this problem, the project develops an adaptive Human-Robot-Interaction system, which enables the robot to read the state (confused, interested, distracted and normal) from students and perform appropriate behaviours as feedback. The whole system is composed of three parts: Facial Expression Analysis (FEA) system, Robotic Behaviour Generator (RBG) and the state-action mappings between them. The FEA system combines cloud computing and supervised classification, without relying on Graphics Processing Units (GPUs). The cloud API takes each frame from the webcam and outputs the intensity values of eight basic emotions. The feature vector containing these values is then classified as a specific state. The second part, the behaviour generator, is designed by the technique of Learn from Demonstration (LfD). In this project, it means to observe the behaviours of the lecturer when the students are showing interest. To collect desired behaviours, the video data from the educational scenarios are analysed by the FEA system. These behaviours serve as the responses to the distracted and confused state. Besides, additional actions are designed to complete the mapping strategies, for example, flashing the LEDs if the students are still distracted. At the end of project, user interface is developed to show the image from the webcam and visualise the current state of students. The follow-up experiments demonstrate that although the proposed system has not excelled other non-interactive systems in improving students' understanding of the content, it has, to a greater extent, enhanced the engagement of students.

Keywords:

Facial emotion recognition, humanoid robot, robotic education

# 摘要

类人机器人已成为提高学生课堂参与度和学习效率的有效工具。但是，在大部分机器人教学场景下，学生的学习状态没有得到重视，导致机器人的作用不能得到最大程度的体现。为解决这个问题，该项目开发了一个自适应的人机交互系统。该系统可以实时监测学生状态（正常，困惑，感兴趣和走神）并予以适当反馈。整个系统由三部分组成：面部表情分析，机器人行为生成器，以及两者之间的映射模型。面部表情分析结合了云计算服务和监督分类算法，能够在不依赖高性能图形处理器（GPU）的条件下运行。云服务提供的 API 能够逐帧读取视频流，并输出 8 种基本情绪的置信度。这 8 个置信度组成的特征向量会被分类为一种特定的状态。第二部分——机器人行为生成利用了模仿学习（Learn from Demonstration, LfD）的设计方法。该项目中，指的是观察实际课堂中学生和讲师的状态，并在学生表现出兴趣时记录讲师的行为。为提取目标行为，面部表情分析系统被用于处理视频数据。这些行为被用于机器人在学生走神，困惑状态时采取的反应。完整的映射模型还包括对同一状态重复时的反应，例如，当学生持续表现走神状态，机器人会闪烁眼部的红灯。同时，为了显示摄像头图像和可视化当前学生状态，该项目实现了一个用户界面。后续的实验证明，相较于无交互的授课模式，本项目提出的系统虽然没有在增强学生对课程内容的理解上体现出优越性，但是它能够在更大的程度上提高学生的课堂参与度。

# Chapter 1: Introduction

Humanoid robotic technology has been bringing new opportunities to the development of education. It is introduced to the STEM (Science, Technology, Engineering, and Mathematics) and language teaching, and proved to improve students' engagement and sustain learning. People also use it to deal with the increasing number of students, limited school budget and personalized teaching.

However, in most cases, the state of the students is not being observed, which may lead to unexpected results: the robot continues teaching even when students are in the negative state. Hence, an adaptive Human-Robot-Interaction (HRI) system is required in robot-aided education.

Intending to identify students' state, this project starts with facial emotion recognition. Some humanoid robots, like the NAO robot, have already embedded with emotion recognition technology. Programmed with the toolkit, the robot can detect five basic emotions (neutral, happy, surprised, angry and sad) from human's facial expression. However, the connection between emotion and state in class lacks theoretical support, even though positive emotions may contribute to high efficiency (Lewis *et al.*, 2010).

To accurately define the state of students, some researchers invite psychoeducational experts to annotate the their behaviours (Saneiro *et al.*, 2014). They associate state with both facial expression and body movement. Some people also use the combination of arousal and valence to describe student's state (Tielman *et al.*, 2014).

Even with a reliable description of the state, the HRI system remains incomplete, since there is a lack of appropriate responses to the state. According to current studies, the robotic behaviours are designed based on theoretical model (Tielman et al., 2014). Although the experiment on students shows the superiority of those designed behaviours, there is no powerful support from data in the real educational scene.

To solve current problems, this project focuses on how facial expression recognition could be adjusted to suit the educational scenario and then implemented on the humanoid robot teacher. The project has finally achieved an adaptive Human-Robot-Interaction system, which can run without GPU and visualise students' current state through GUI. It consists of Facial Expression Analysis (FEA) system, Robotic Behaviour Generator (RBG) and the state-action mappings between them.

First, the project adopts Microsoft Face API to realise facial emotion recognition. Facial expressions of students are read from webcam and stored as frames. For each face in a frame, the API returns intensity values of 8 basic emotions. Then, to further describe the state of students in the class, the project proposes a new model. Students are invited to perform posed expressions when they are confused, interested, distracted and in normal state. These labelled expressions are used in the training phase of the K-Nearest Neighbours (KNN) classifier. The trained classifier will categorise different combinations of the intensity values into four states.

Next, for designing appropriate response to different state, the project observes behaviours of the lecturer in the real educational scene. Those behaviours that trigger the positive state from the students are collected, as the response to confused and distracted state. Besides postures and gestures, the state-action mappings also involve the change of LEDs and volume. The robot will perform this kind of action if the students are still distracted or confused.

While implementing the desired behaviours, the robot's SDK is used. Since the SDK only supports Python 2, the RBG system is completed with Python 2. The connection between FEA system and RBG system is a socket, which operates in the subscribe-publish pattern.

After connecting to the integrated HRI system, the robot teacher will detect the state of the students while delivering a lecture. The real-time image will be demonstrated on the GUI, with the boxes localising the faces. The current state of each face can be read from the box's colour. There is also a horizontal bar chart showing the proportion of students in each state. The state of multiple students is defined by the state with the highest proportion. If the negative state, like distracted or confused, is identified by the robot, it will perform corresponding behaviours. At the end of the research, the HRI system is evaluated in a simulated teaching scenario. Participants were invited to the experiment and fill up the questionnaire. Although the proposed system does not show superiority in increasing students' understanding, it is proved to improve the engagement of the students to a greater extent.

The rest of the report is organized as follows: Chapter 2 presents an overview of robot-aided education, technology related to facial expression analysis, and the design of robotic behaviours. Chapter 3 introduces the methodology of the project. The evaluation process and discussion of results are presented in Chapter 4. Conclusion and future work are provided in the last chapter.

# Chapter 2: Background

## 2.1 Robot-aided education

### 2.1.1 Existing robotics education

With the evolution of robotic technology and the need of quality enhancement in teaching, more and more robots are incorporated into education. These educational robots get involved in similar activities as the teacher, tutor or peer confidence (Belpaeme *et al.*, 2018). As a teacher, the robot delivers lectures directly, with its pre-determined gestures and teaching materials. As a tutor, it cooperates with the human teacher, improving the engagement of the learners. As a learning companion, it would act as a novice. In this case, the value of the robot is to boost learner's (see Figure 1).

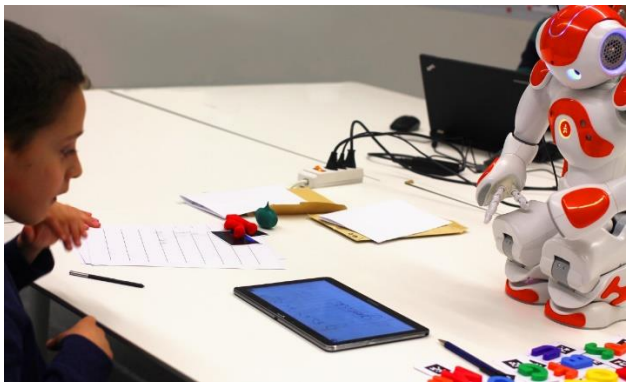


**Figure 1. Robot help with the handwriting acquisition (CHILI, 2018)**

Regardless of which role it performs, the robot is proved to arise the students' interest in various fields, such as language, programming and so forth.

The language teaching robot is commonly used in the education of early childhood. In Europe, a research project called L2TOR has been launched. This project aims to use the NAO robot for preschool students to study a second language (LTUTOR, 2018).

Also, in some middle and high school, the robot has become popular as a teaching tool. Since most students are novice programmers, simple robotic programming is suitable for serving as an introductory course. Robot kits like Lego Mindstorms (see Figure 2) and Robotis Bioloid are available for students to control the robots, helping them learn about science, technology, engineering and mathematics (STEM) knowledge.

**Figure 2. LEGO EV3RSTORM (LEGO, 2013)**

### 2.1.2 Interaction enhancement

Instead of STEM and language education, this research shifts the emphasis to human-robot interaction (HRI), that is, to make the robot teacher more human-like and further cover the shortage of human teacher. If a robot has the ability to notice the state of students, it could adapt its behaviours accordingly to improve its teaching quality.

One practical approach is to consider the emotional status of being. This is also an important issue considered in affective HRI system. Some researchers hold the belief that there exists a interplay between emotion and learning-related cognitive processes (Lewis et al., 2010). A positive mood leads to better performance in thinking and decision making while negative or natural mood induces impairment of learning efficiency.

Since emotion is expressed through both verbal and nonverbal ways, the methods for detecting are various. In some studies, tiny fluctuation of emotion is observed by biometric tools: with electroencephalography (EEG), which shows the electrical activity of the brain; or the Galvanic Skin Response (GSR), which senses body temperature and perspiration. Besides those unconscious reactions, gesture analysis is also used in observing emotion variation by the body movements.

With the development of machine learning, more solutions are available for automatic emotion recognition system. For example, voice recognition and facial expression recognition.

## 2.2 General process of facial emotion recognition

Compared to other biometric methods, the facial expression is much easier to obtain when it comes to the teaching scenario. An emotion recognition process generally consists of three

phases: face detection, feature extraction, and emotion classification.

### 2.2.1 Face detection

The goal of face detection is to locate the faces in the image, by distinguishing the face areas from non-face ones.

Skin colour-based detection is a common and effective technique (Elgammal *et al.,* 2009), especially when the background colour is controlled. To identify face regions in the image, pictures are first transformed to a suitable colour space. As is often the case with face detection, HSV (Hue, Saturation, Value) colour model advantages over the RGB (Red, Green, Blue) colour model. Then, pixels are classified as face and non-face colour. The classifier applied is trained by a database of skin-coloured patches from different images.

Viola-Jones framework was proposed by Paul Viola and Michel Jones (Viola and Jones, 2001). Instead of working directly with image intensities, the framework extracts Haar-like features from a new image representation called integral image. The computation of the integral image requires only a few operations per pixel so that the feature evaluation could happen in constant time. To effectively determine face area from a large set, a modified AdaBoost algorithm and a cascaded classifier are also applied. OpenCV, a computer vision library, has provided pre-trained Haar feature-based cascade classifiers for face detection.

With the development of deep learning, researchers present techniques based on Deep Neural Network (DNN) to improve accuracy and speed. Single Shot Multibox Detector (SSD) is an emerging framework for object detection (Liu *et al.,* 2016). Rather than processing the whole image, it uses bounding boxes to locate possible areas where the objects exist. Then, through the small convolutional filters applied to feature maps, category scores and box offsets of those bounding boxes are predicted. Also, this framework can be implemented with the OpenCV library.

### 2.2.2 Feature extraction

Feature extraction aims to figure out distinctive features from the detected faces. According to the present study, a general extracting framework consists of four major parts: filtering, encoding, spatial pooling and holistic representation (Wang *et al.,* 2018).

To begin with, the image is convolved with a specific filter to get the local features, which are then encoded to a histogram or a feature vector. Next, to generate a global representation, the

following spatial pooling is implemented. Once the pooling phase has finished, the holistic encoding outputs the final feature vector of the previous local features within the whole image.

Local Binary Patterns (LBP) is a simple but effective technique for feature extraction (Ojala *et al.*, 1994). The main idea is to compare each pixel with adjacent ones, and then store the results in binary form. Because of its simplicity, it is suitable for real-time analysis of images.

### 2.2.3 Emotion classification

Emotion classification matches the expression to a specific emotion based on extracted features (eyebrows, eyes, and mouth). In general, researchers develop their classification models in categorical approach or dimensional approach (Grekow, 2018).

The categorical approach describes emotion discretely. There are four current models of basic emotions (see Table 1, the dashed line indicates the controversial emotion marked by the author) proposed by Ekman and Cordaro (see Figure 3), Izard, Levenson, and Panksepp and Watt.

**Table 1: Four models of basic emotions**

| Ekman & Cordaro | Izard | Levenson | Panksepp & Watt |
|---|---|---|---|
| Happiness | Happiness | Enjoyment | Play |
| Sadness | Sadness | Sadness | Panic/Grief |
| Fear | Fear | Fear | Fear |
| Anger | Anger | Anger | Rage |
| Disgust | Disgust | Disgust | Seeking |
| Contempt | Interest | Interest | Lust |
| Surprise | Contempt | Love | Care |
| | | Relief | |


**Figure 3. Basic emotions (Lucey *et al.*, 2010)**

These models have a great similarity. The positive emotion is represented by Happiness (Ekman and Cordaro; Izard), enjoyment (Levenson) and Play (Panksepp and Watt). And they all contain three distinct negative emotions: sadness (defined as Panic in Panksepp and Watt's model), fear,

and anger (Tracy and Randles, 2011).

As for the dimensional approach, it identifies emotion by its location in a space of numerical dimensions, for example, the PAD emotional state model. It deals with the perceptions of physical environments in three nearly orthogonal dimensions: Pleasure, Arousal and Dominance (see Figure 4). Pleasure measures the extent of enjoyment, while the Arousal reflects how large the environment excites and simulates the individual. The Dominance deals with how much that individual feels being in control.

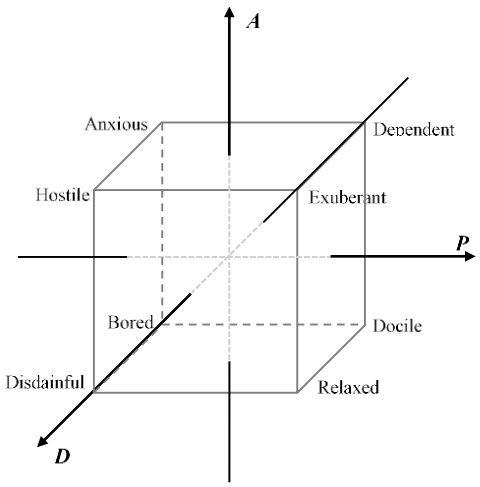



**Figure 4. The PAD model**

By reference to the emotion model, the expression is classified by computing the similarity of it and labelled faces. The distance is a commonly used metric. For instance, the Euclidean distance based method calculates the distance between the final score weighted matrixes of faces. In this case, the labelled image with the lowest value is chosen as the result. The Line segment Hausdorff Distance (LHD) method compares two sets of line segments within the edge maps of faces. The edge map refers to the line segment representation of the object (Gao and Leung, 2002). The Minimum Distance Classifier (MDC) measures the distance between the feature vectors in sub-images containing left eye, right eye, mouth, etc. (Revina and Emmanuel, 2018).

## 2.3 State-of-art emotion recognition API

At present, there are already well-developed tools which integrate three phases of emotion recognition, providing users the access to the functions by APIs or SDKs.

### 2.3.1 Microsoft Face API

Microsoft Face API is a cognitive service that supports detection, recognition, and analysis of human faces in images (Microsoft, 2019). Once the image is passed to the API, the detected faces will be stored with unique face IDs temporarily for 24 hours. During the analysis, the API locates face components (i.e., eyes, nose, mouth and eyebrows) with 27 facial landmarks. Then, it would identify several attributes for the face, including facial emotion.

The emotion is represented by the intensity of 8 basic emotions (Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness and Surprise), with reference to the work of Ekman and Cordaro. Each intensity is scored from 0 to 1. In addition to expression analysis, the API also supports identifying the similarity of two faces. The similarity confidence of candidate faces ranges between [0,1].

### 2.3.2 Face++ API

Face++ (Megvii, 2019) also offers face recognition for images. A little different from the Face API, it considers 7 basic emotions (Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise). The value of each emotion is between [0, 100]. As for locating, it provides up to 106 high-precision facial key points.

Besides face recognition, Face++ has body recognition as well. By passing image file to the Skeleton Detect API, it would feedback with 12 key points of each detected body.

### 2.3.3 Google Vision API

Google Vision API (Google, 2019) is another commonly used analysis tool. Unlike Microsoft Face API and Face++ API, the Vision API only considers 4 emotions: joy, sorrow, anger, surprise. Each emotion is evaluated in 5 degrees (except 'Unknown'): Very Unlikely, Unlikely, Possible, Likely and Very Likely. It locates each face by 34 three-dimensional landmarks.

## 2.4 Supervised classification algorithm

### 2.4.1 K-Nearest Neighbours

K-Nearest Neighbours (KNN) is a supervised classification technique, which means the classifier is trained by labelled samples. Each sample has feature values and the label of a specific class. The object would be classified by its feature values. As illustrated in Figure 5, the main idea is to mark K closest samples to the object and then find the class that most samples

belong to. Here, the distance is calculated using feature values (Scikit-learn, 2018).

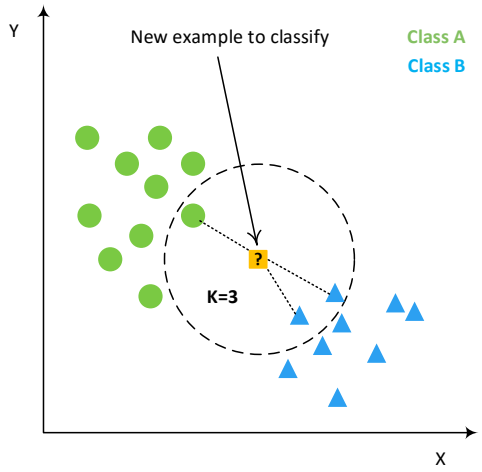



**Figure 5. KNN algorithm**

While finding the nearest neighbours, the selection of the algorithm depends on the size and dimension of the samples. The brute-force algorithm is the simplest but inefficient in a large number of data. The K-D Tree shows good performance for a large dataset, but impractical in high dimension ($D > 20$). The Ball-tree algorithm is invented for computation in high dimensional space. However, the execution would be slower than the K-D Tree in low dimension ($D \leq 20$).

### 2.4.2 Linear Support Vector Machine (LSVM)

The Support Vector Machine (SVM) is a supervised algorithm. The main idea is defining a hyperplane that best divides samples into two classes (see Figure 6). Based on the SVM, the linear SVM algorithm is proposed for the classification of multiclass and high dimensional data.

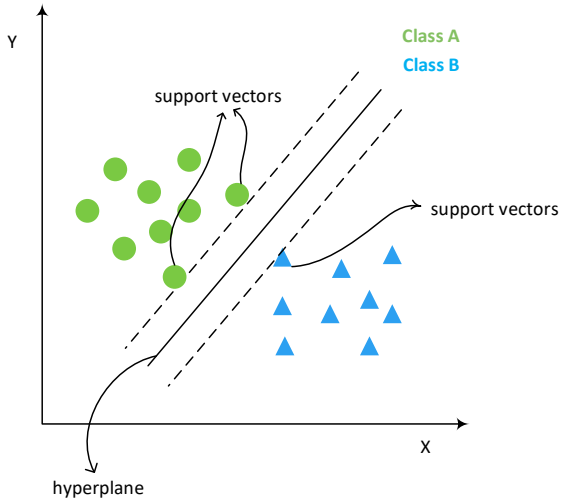



**Figure 6. SVM algorithm**

### 2.4.3 Neural Networks

The neural networks can be used in supervised classification. It includes three parts: input layer, hidden layer, and output layer. Each layer is composed of a series of neurons. The hidden layer often has multiple layers. A neuron computes the input values and outputs activation function, which defines the class that the object belongs to (see Figure 7).

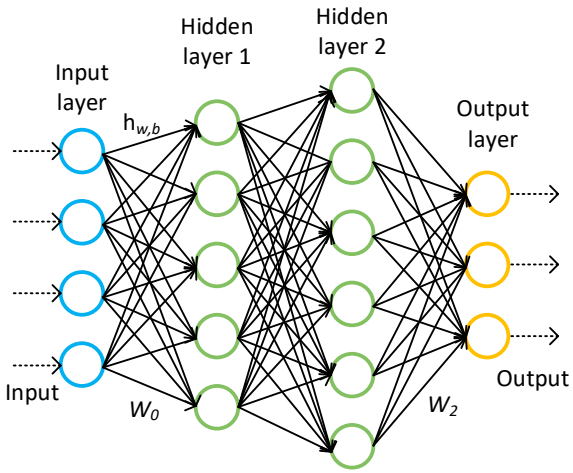



**Figure 7. Neural networks**

## 2.5 Robotic behaviour

### 2.5.1 Existing emotion models

Similar to emotion classification, many researchers design their robotic behaviours by referring to basic emotion models or dimensional emotion models. As for the basic emotions, they use a simple pattern of each emotion to express emotions (Liu *et al.*, 2017). In the dimensional approach, they use weighted values for each dimension to design expressions (Breazeal, 2003).

### 2.5.2 Learn from Demonstration (LfD)

Instead of using existing models, the robotic behaviours could be also learned from human's demonstration.

Learning from Demonstration, also known as "Imitation Learning" and "Teaching by Showing", is a technique for designing robotic behaviours from the observations of human's performance (Argall *et al.*, 2009). It could, as a result, provide mapping policies between current world state and robotic actions. The whole process generally consists of two phases; one is collecting demonstration examples; the other is deriving a mapping policy from the examples.

During the collecting process, the data is recorded as state-action pairs from human activities.

The source varies from sensors of robot teleoperated by the human, to cameras recording a human's behaviours. The methods are categorized by demonstration and imitation. Demonstration approach is implemented on the actual robot, while imitation means the behaviours are not performed directly on the robot, but through the sensors.

To derive policies from the dataset, there are three core approaches: mapping function, system model and plan. The mapping function is a function that approximates the mapping between state and action. The system model is a state transition model based on the structure of Reinforcement Learning (RL). The plan represents the policy as a series of actions. These actions are defined as the pre-conditions, the state before the action performing, and the post-conditions, the following state after the execution.

## 2.6 The Pepper robot

Pepper is a humanoid robot equipped with the ability of emotion detection and voice analysis. It has currently been used in various application scenarios, such as offices, schools, and homes (Softbank, 2019).

### 2.6.1 Kinematics data

Pepper is composed of links, joints and body frames (see Figure 8). It has one head, two arms and a single leg that are all connected by the torso. The leg is composed of thighs and base, with three omnidirectional wheels at the bottom. Pepper's leg and torso provide with 3 Degree of Freedom (DoF). Its head has 2 DoF at the neck, and each arm has 5 DoF at the joints. The base part has 3 DoF because of the wheels.

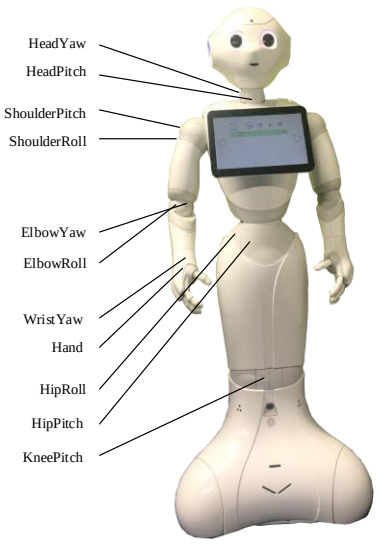



**Figure 8. The Pepper robot**

### 2.6.2 NAOqi SDK

NAOqi is the robot kits for the Pepper. It supports programming languages like Python, C++ and etc. For Python, it only supports Python 2. According to the documentation, the NAOqi SDKs provide with the functions of controlling the robot in motion, audio, people perception, sensors, and LEDs.

### 2.6.3 Choregraphe suite

Choregraphe suite (see Figure 9) integrates the NAOqi SDKs to a desktop application, making the control of the robot easier. There are a series of pre-defined functions boxes. Each box is loaded with Python code. Dragging the boxes and connecting them, the code would be loaded and executed to control the robot. The code inside the box is available to be modified. Also, it is able to simulate the designs by connecting to the virtual robot. With the simulation, the design process becomes efficient and effective.

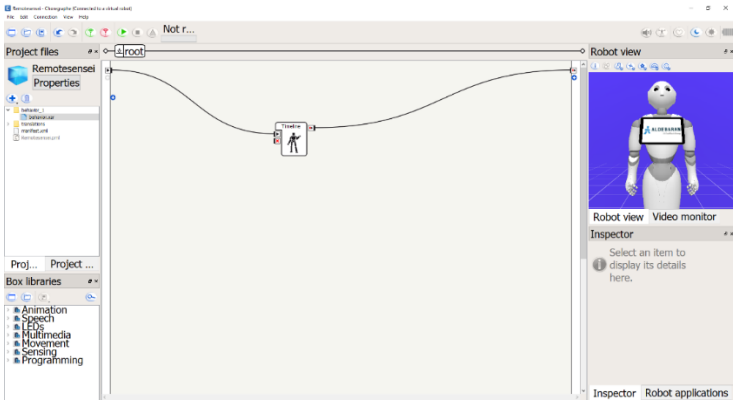



**Figure 9. Choregraphe GUI**

## 2.7 Discussion

Among the existing facial expression database, this project selects the CK+ dataset, since it provides the sequence ranges from neutral to the extreme expression, which is appropriate for evaluating the expression analysis tool. As for the facial expression analysis tool, the Microsoft Face API is selected because it refers to the same Emotion model as the CK+ dataset. To further classify the expression by the results of facial expression analysis, the project follows the instructions on the official website of Scikit-learn. Since the training of Neural Network requires a large training set size and the processing of GPU, it is not practical for this project. With reference to the instructions, the project finally selects KNN algorithm, since it shows better performance than the linear SVM in accuracy. To classify data with 8 features, the K-D Tree algorithm is used for finding neighbours. In addition, on the graphical user interface, each

face displayed is located with a rectangular. Here, the face detection is realised by OpenCV-Python library. Instead of typical Haar feature-based cascade classifiers, the frames are processed by the SSD framework, which provides better accuracy.

# Chapter 3: Design and Implementation

## 3.1 System architecture

The HRI system consists of two main modules: the facial expression analysis (FEA) system and the robotic behaviour generator (see Figure 10). They are related through state-action mappings. After connecting to the HRI system, a robot teacher is able to detect students' state while delivering a lecture. To start with, the real-time data from webcam enters the FEA module. The emotion recognition gives intensity value, the component in a feature vector, for each basic emotion. Then, the feature vector is classified as a description of current students' state. Based on the mapping strategy (which is explained in section 3.4), the state is matched to the corresponding behaviour. The generator will look up the control code of that behaviour and send it to the robot if any actions are needed.

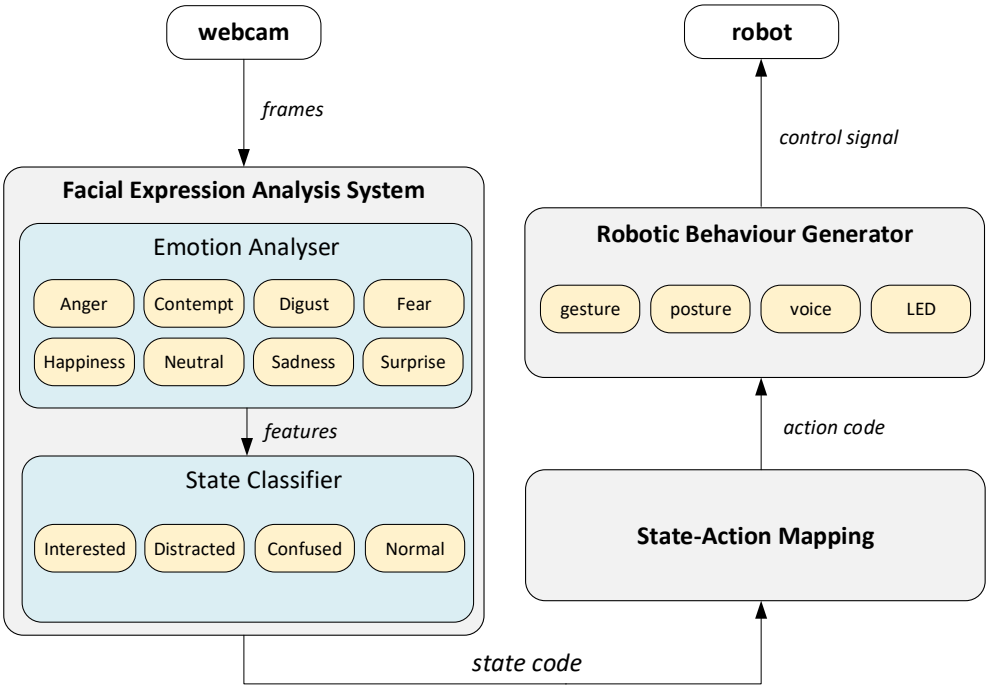



**Figure 10. System architecture**

## 3.2 Dataset construction

### 3.2.1 Image dataset

The image dataset would be used in the construction of the FEA module. It is composed by an existing facial expression dataset and posed behaviours of students.

Many well-annotated facial expression databases have been published for the research of

expression recognition. The datasets, such as CK+, JAFFE, and FERG, contain facial expressions with a range of basic emotions. In this research, CK+ dataset was selected. It extends from the CK dataset, involving 7 basic emotions described in the Ekman and Cordaro's emotion model. There are 593 sequences across 123 subjects, while 327 of them are labelled with emotion. For each emotion, the participants are asked to present posed expressions with the increment of intensity. Take the first sequence (see Figure 11) as an example, the face changes from neutral to the extreme disgust, which could be appropriate for evaluating the expression analysis tool.

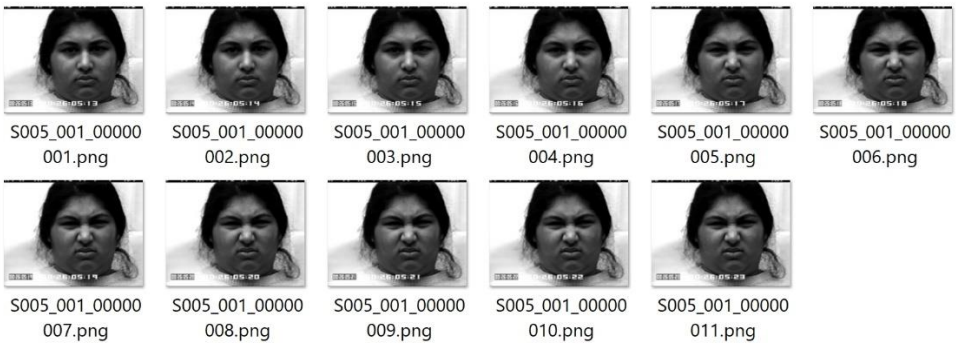



**Figure 11. The first sequence in CK+ database**

Another image dataset was built to refine the analysis system (see section 3.3.3 for more details). The students were invited to show how they usually behaved in class. They performed posed expressions when they were confused, thinking, distracted and in normal state. For each state, 50 samples were collected. Figure 12 shows part of the dataset.

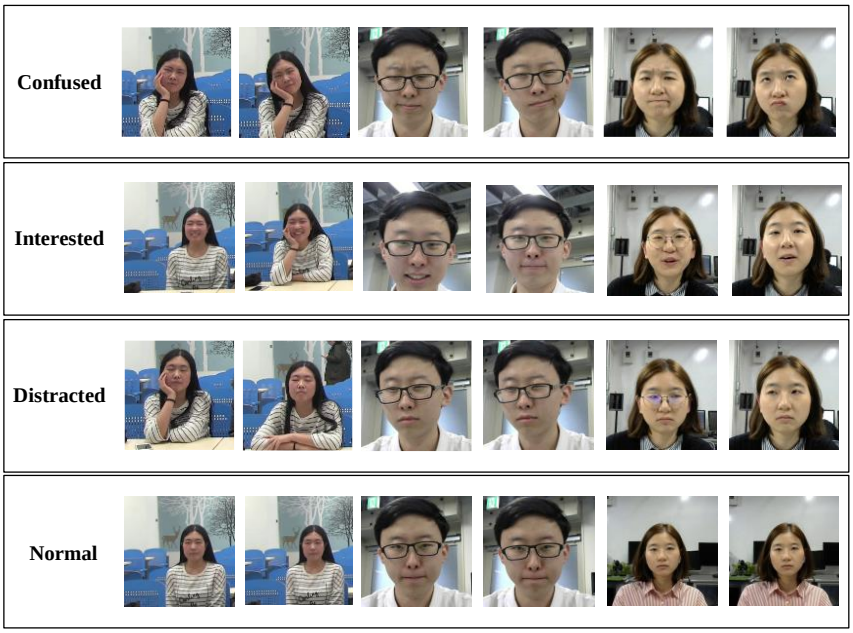



**Figure 12. Part of posed behaviours**

### 3.2.2 Video dataset

The video data were collected in both real lectures and simulated scenes.

The videos of real lectures were taken in university classes. To identify desired behaviours that could keep students in the positive state, the project settled two cameras, one was face to the students, and the other was face to the lecturer. In this case, behaviours of both students and teacher were recorded.

For the simulated one, several students (2 to 4) were invited to watch videos and their reactions were recorded. The videos were selected from speeches and debates, which approached the teaching scenario. In this process, only one camera was used. In addition, to easily synchronize the videos during data processing, the volume of videos was adjusted high enough to make sure it was clearly recorded by the camera.

## 3.3 Facial expression analysis (FEA)

### 3.3.1 Implementation

Since the Microsoft Face API refers to the same emotion model (Ekman and Cordaro's work) as the CK+ dataset, this research implements facial emotion analysis by this tool, in Python 3. The image is first transformed to binary form, and then passed to the API by Requests Python library. After requesting, the Face ID, emotion intensities and facial landmarks for each face will be sent back in JSON form (see Figure 13).


```
[
    {
    "faceId":
    "48e30983-18e9-4f4f-8011-aea3ecfc3d02",
    "faceRectangle": {
       "top": 128,
       "left": 459,
       "width": 224,
       "height": 224
     },
     "faceAttributes": {
     "emotion": {
       "anger": 0.0,
       "contempt": 0.0,
       "disgust": 0.0,
       "fear": 0.0,
       "happiness": 1.0,
       "neutral": 0.0,
       "sadness": 0.0,
       "surprise": 0.0
       },
     "faceLandmarks": {
       "pupilLeft": {
         "x": 504.8,
         "y": 206.8
       },
       "pupilRight": {
         "x": 602.5,
         "y": 178.4
       },
       "eyeLeftOuter": {
         "x": 490.9,
         "y": 209.0
       },
       "eyeLeftTop": {
         "x": 509.1,
         "y": 199.5
       },
       "eyeLeftInner": {
         "x": 529.0,
         "y": 205.0
       },
       "eyeRightInner": {
         "x": 590.5,
         "y": 184.5
       },
       "eyeRightOuter": {
         "x": 623.8,
         "y": 173.7
       },
                                    ... ]
```


**Figure 13. Response of API**

As the Face API only takes binary image data as input, the videos are split into frames and then converted to binary streams. However, according to the mechanism of the Face API, it assigns

different and unique IDs to detected faces. Meanwhile, the Face IDs are sorted randomly, without regard to the faces' locations. In this way, the ID of the same person in different frames will be inconsistent, which disables the face tracking.

To deal with this problem, the project considers the face verification service of the Face API. In the beginning, the system gets the first frame and stores the Face IDs in a list. The Face IDs are rearranged from left to right, by horizontal coordinate extracted from the "left" of the "faceRectangle" attribute. Then, for the following frames, each new ID received is compared with IDs in the list (see Figure 14).

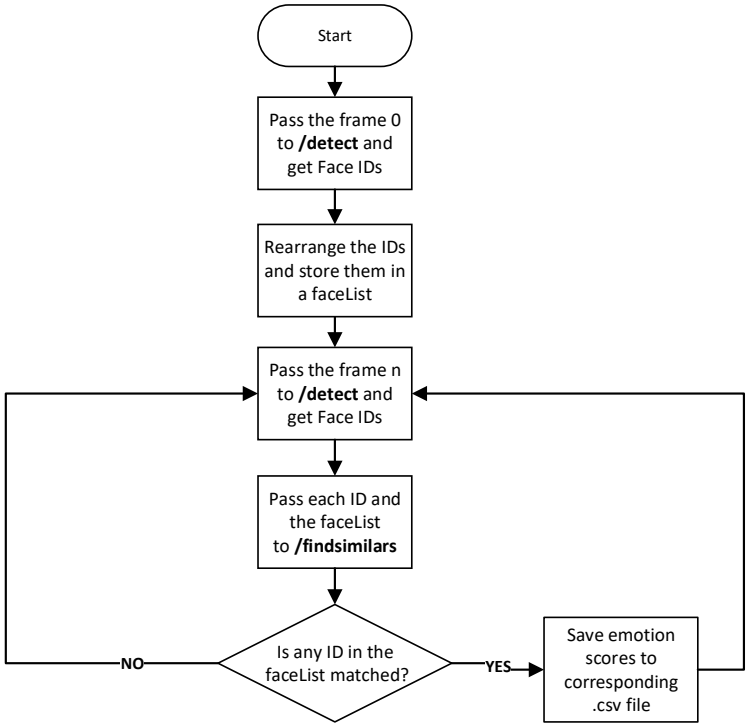



**Figure 14. Flow chart of real-time detection**

In the case of teaching scenario, the appearances of students are less likely to be changed, so that the confidence score of the similarity could be guaranteed. Meanwhile, the seats of students are fixed, which makes their localisation possible. Therefore, the improved FEA system is able to track each student in a group emotion recognition.

### 3.3.2 Experiment I: Evaluation of the FEA system

This experiment was first conducted on the CK+ database. Since the database was partially labelled, the experiment selected 7 sequences labelled with 7 different emotions. As can be seen from the Table 2, the results from a sequence of sad expressions (see Figure 15), only when the expression was obvious enough (i.e., at level 6) could the system correctly identify the emotion.

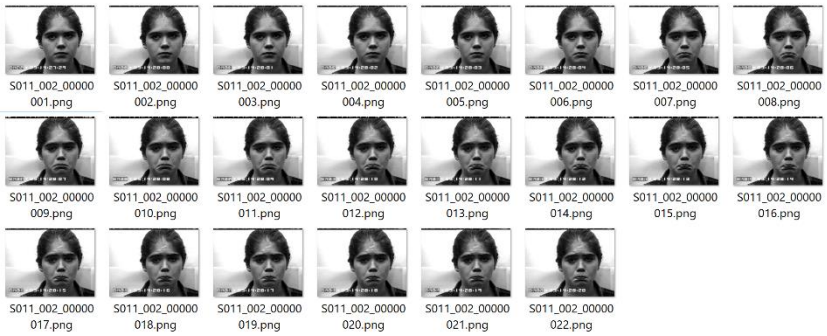



**Figure 15. Sad expressions**

**Table 2: Results of sequence S011_002**

| | anger | contempt | disgust | fear | happiness | neutral | sadness | surprise |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0.999 | 0.001 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0.999 | 0.001 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0.999 | 0.001 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0.996 | 0.004 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0.988 | 0.011 | 0 |
| 5 | 0 | 0.001 | 0 | 0 | 0 | 0.939 | 0.061 | 0 |
| 6 | 0 | 0.005 | 0 | 0 | 0 | 0.258 | 0.737 | 0 |
| 7 | 0 | 0.004 | 0 | 0 | 0 | 0.032 | 0.964 | 0 |
| 8 | 0 | 0.002 | 0 | 0 | 0 | 0.002 | 0.996 | 0 |
| 9 | 0 | 0.001 | 0 | 0 | 0 | 0.002 | 0.997 | 0 |
| 10 | 0 | 0.001 | 0 | 0 | 0 | 0.001 | 0.998 | 0 |
| 11 | 0 | 0.001 | 0 | 0 | 0 | 0.003 | 0.996 | 0 |
| 12 | 0 | 0.002 | 0 | 0 | 0 | 0.019 | 0.979 | 0 |
| 13 | 0 | 0.001 | 0 | 0 | 0 | 0.004 | 0.995 | 0 |
| 14 | 0 | 0.001 | 0 | 0 | 0 | 0.006 | 0.993 | 0 |
| 15 | 0 | 0.001 | 0 | 0 | 0 | 0.009 | 0.99 | 0 |
| 16 | 0 | 0.001 | 0 | 0 | 0 | 0.039 | 0.96 | 0 |
| 17 | 0 | 0.002 | 0 | 0 | 0 | 0.035 | 0.963 | 0 |
| 18 | 0 | 0.001 | 0 | 0 | 0 | 0.098 | 0.9 | 0 |
| 19 | 0 | 0.001 | 0 | 0 | 0 | 0.079 | 0.92 | 0 |
| 20 | 0 | 0.001 | 0 | 0 | 0 | 0.05 | 0.95 | 0 |
| 21 | 0 | 0.004 | 0 | 0 | 0 | 0.058 | 0.937 | 0 |

In addition, the system showed unexpected performance in differentiating fear and sadness (see Table 3 and, Figure 16).


**Figure 16. Fear expressions**

**Table 3: Results of sequence S011_003**

| | anger | contempt | disgust | fear | happiness | neutral | sadness | surprise |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0.999 | 0.001 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0.999 | 0.001 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0.999 | 0.001 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0.998 | 0.002 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0.998 | 0.002 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0.994 | 0.006 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0.974 | 0.026 | 0 |
| 7 | 0 | 0.001 | 0 | 0.001 | 0 | 0.647 | 0.351 | 0 |
| 8 | 0 | 0 | 0 | 0.007 | 0 | 0.289 | 0.701 | 0.002 |
| 9 | 0 | 0 | 0 | 0.03 | 0 | 0.074 | 0.894 | 0.001 |
| 10 | 0 | 0 | 0 | 0.042 | 0 | 0.085 | 0.869 | 0.003 |
| 11 | 0 | 0 | 0 | 0.008 | 0 | 0.016 | 0.976 | 0 |
| 12 | 0 | 0 | 0 | 0.015 | 0 | 0.017 | 0.967 | 0 |
| 13 | 0 | 0 | 0 | 0.009 | 0 | 0.012 | 0.978 | 0 |

Also, the experiment with real-time webcam revealed drawbacks of the system. When the expression was not obvious, the expression was detected as “neutral”.

Besides, from the theoretical point, although researchers have proved the relationships between positive emotions and high efficiency, basic emotions are insufficient to describe students’ state during the lecture.

### 3.3.3 Refinement

To overcome problems in the previous system, this project attempts to describe the expression in a higher level: interested, distracted, confused and normal.

In this section, the second image dataset is used. Images are grouped according to the 4 states, and then processed by the Face API. Given 8 emotion intensities as features and 4 states as labels, the system is constructed using supervised classification techniques (see Figure 17).

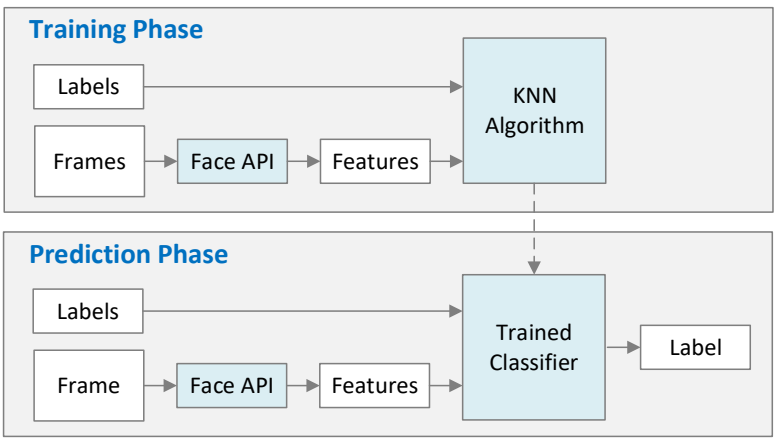



**Figure 17. Training process and prediction process**

During data analysis, a vague boundary is found to exist between patterns of the "normal" and the "distracted" (see Figure 18). Hence, a new metric, eye movement, is introduced to differentiate these two states.

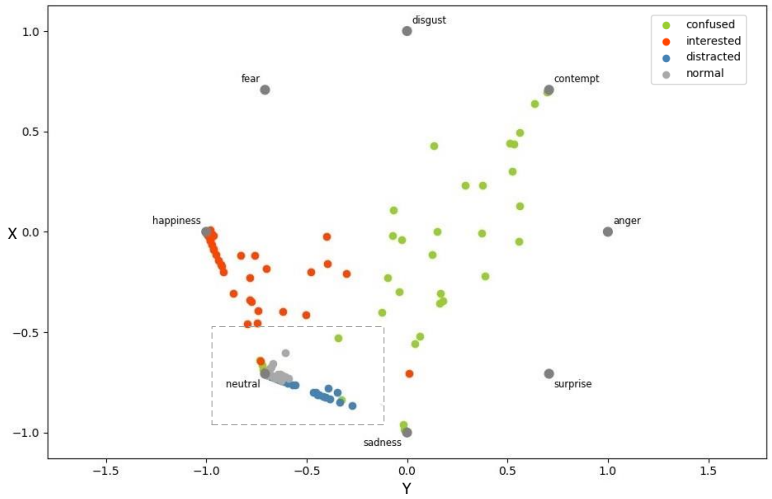
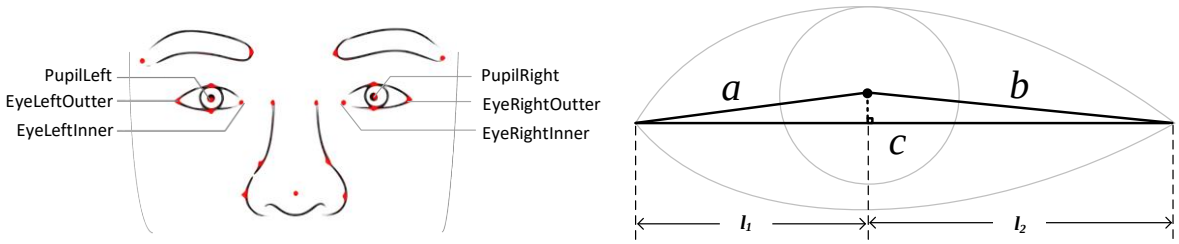



**Figure 18. Pattern of four states**

According to the facial landmarks return by the Face API (see Figure 19a), the locations of pupils and orbits are included, which make it possible to detect eyes' distraction from the front side. The calculation extracts the coordinates of three points: the pupil, the inner point, and the outer one. As shown in Figure 19b, line segments $a$, $b$, and $c$ are the Euclidean distances between each pair of these points. $l_1$ is the projection of $\boldsymbol{a}$ onto $\boldsymbol{c}$, while $l_2$ is the projection of $\boldsymbol{b}$ onto $\boldsymbol{c}$. To figure out the projections, the Law of Cosines is used:

$$l_1 = a\cos B = \frac{a^2 + c^2 - b^2}{2c}, \qquad l_2 = b\cos A = \frac{b^2 + c^2 - a^2}{2c} \tag{1}$$

$$ratio = \frac{l_1}{l_2} = \frac{a^2 + c^2 - b^2}{b^2 + c^2 - a^2} \tag{2}$$




**(a)** **(b)**

**Figure 19. Calculation of eye movement**

The following experiments showed that the ratio ranged in [0.4, 1.5] when people were looking straight. That means, if the expression is classified as normal, it will be re-classified as distracted when the ratio is out of [0.4, 1.5].

To classify data into four states, Scikit-learn, a tool for machine learning in Python, is used. The selection of the classification algorithm is based on the instructions provided by its official website (see Figure 20, and see Appendix A for the whole map).

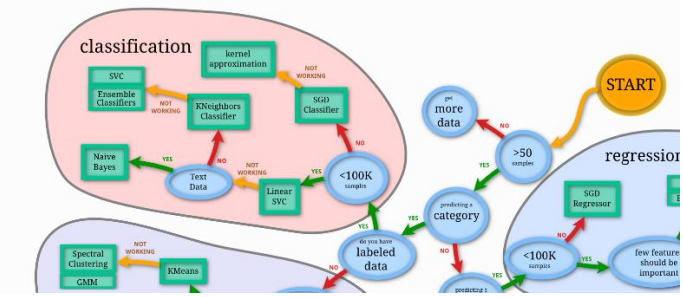



**Figure 20. Machine learning map (scikit-learn official website, 2019)**

Since the dataset is of relatively small size and high dimension, the selection starts with the Linear SVC (LSVM). The parameter C in LSVM classifier refers to the "cost" of prediction, which determines the influence of misclassification. The larger C value provides higher accuracy but may lead to over-fitting. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, reducing the accuracy. To figure out the optimal C value, this project applied a technique called k-fold cross-validation. The labelled data are randomly divided into k equal sized groups. The first group is used for validation while other k-1 groups is for training. The k is set to a commonly used value10. According to the result shows in Figure 21, the optimal C value approaches 21.

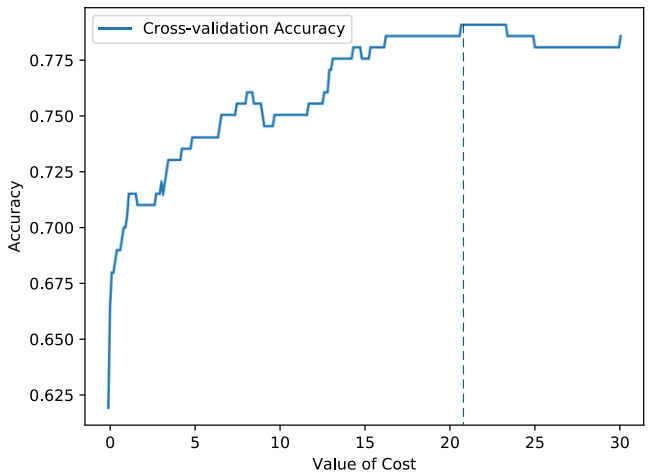



**Figure 21. Cross-validation of LSVM**

However, as can be seen from the diagonal of its confusion matrix, the accuracy of classifier is lower than expected (see Figure 22).

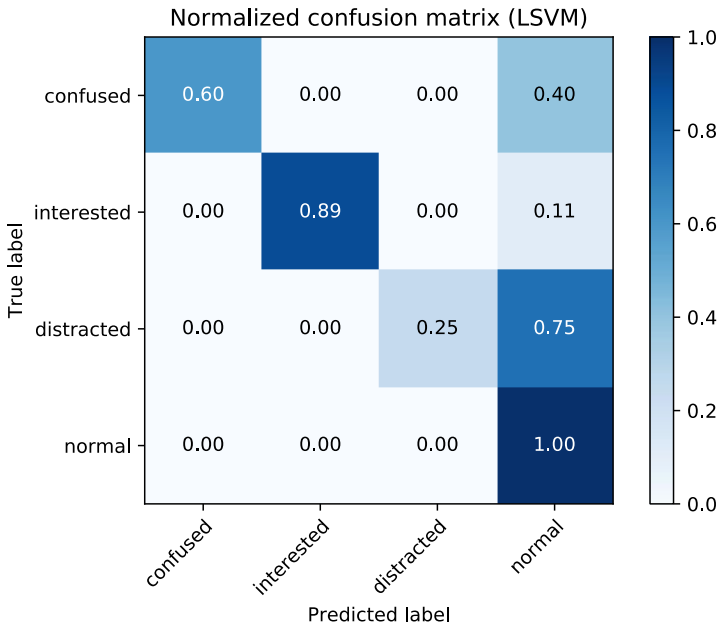



**Figure 22. Confusion matrix of LSVM**

Then, to improve the performance, the next candidate algorithm, KNN, was considered.

The selection of parameter K in the KNN classifier followed the same procedure (see Figure 23). The optimal K value is 4.

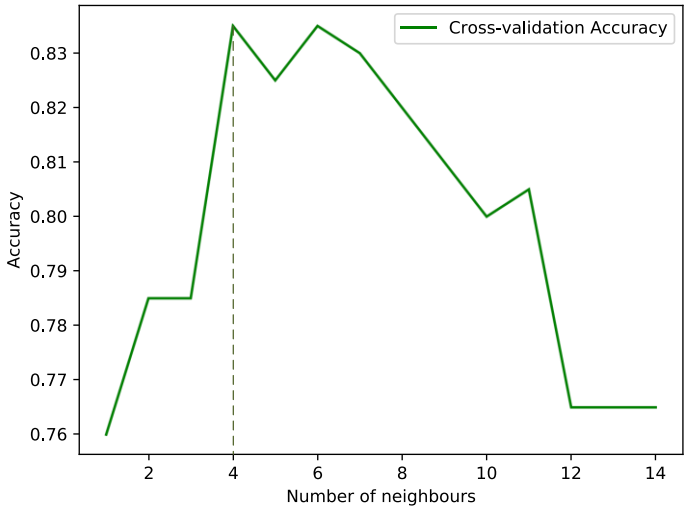



**Figure 23. Cross-validation of KNN**

Its confusion matrix shows a better accuracy (see Figure 24).

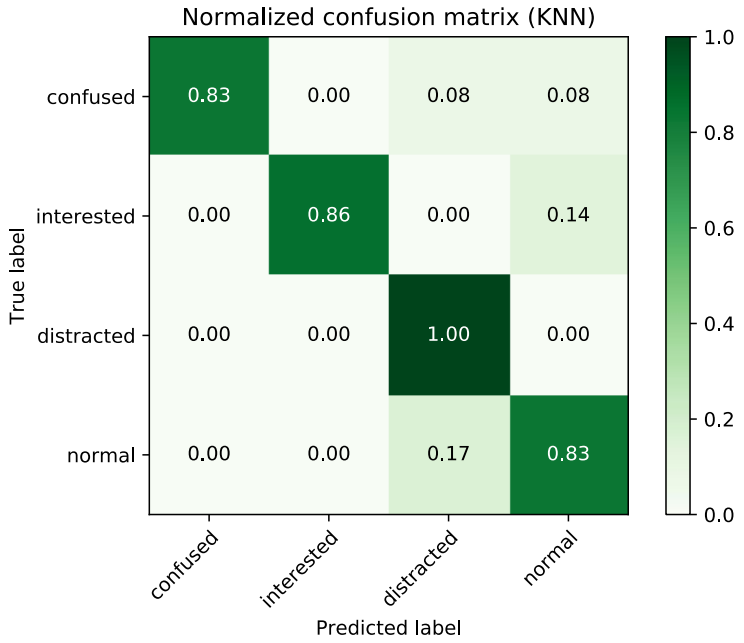



**Figure 24. Confusion matrix of KNN**

After the improvement, the system could identify the state from the facial expression. The system first analyses the basic emotions from the expression and then classifies it as a specific state. If the state turns out to be “distracted” or “normal”, the system carries out an additional computation of eye movement. When detected eyes are not keeping on the side of the robot teacher, the state is classified as “distracted”.

To define the resulting state used in state-action mapping, the project applies a sliding window. More specifically, the state is defined when it keeps the same three times, which is approximately the duration of the students’ positive state (see Experiment II in section 3.4.1). In addition, if multiple people are detected, the state will be decided by the majority.

### 3.3.4 Graphical User Interface

A graphical user interface is developed to show the webcam image and visualise the current state of each student (see Figure 25). It is built with PyQt5 Python library. When the “Capture Start” button is pressed, the images captured by the webcam will be presented. The current states are visualised by a horizontal bar chart, in different colours. Also, faces in different state are located with boxes in corresponding colours.

The localisation is implemented by the SSD framework of OpenCV, which is faster and more accurate than the Haar feature-based cascade classifier. However, the facial expression analysis

is time-consuming, which causes a long delay in execution. For further improvement, the program is modified by decoupling UI from the logic of face recognition, using multithreading.

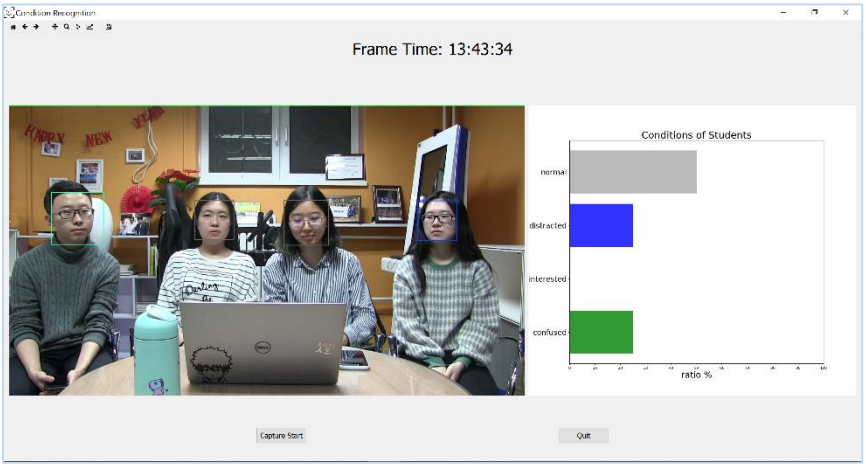



**Figure 25. Graphical User Interface**

## 3.4 State-action mapping

### 3.4.1 Experiment II: Relating the students' state to the lecturer's behaviour

To give an appropriate response to each state, mapping strategies are needed.

The video dataset was analysed to find the relationship between the state of students and the behaviour of the lecturer. The idea is to figure out the behaviour that could improve the engagement of students. Before starting, videos from students and the lecturer were synchronised by the voice recorded from each other.

To begin with, facial expression analysis was carried out. During the period where the majority of students were showing interested, the corresponding behaviour of the lecture was collected. To mark the behaviour, the time stamp of each frame was calculated by multiplying the frame rate by the frame number.

In this way, a series of desired behaviours were recorded in video clips, as the robot's response to the "distracted" state.

### 3.4.2 Mapping strategies

The robot cannot response to each student but to the majority of the group. When most students are showing "distracted" or "confused", corresponding actions are taken, to bring the state back to "interested" or "normal" (see Figure 26).

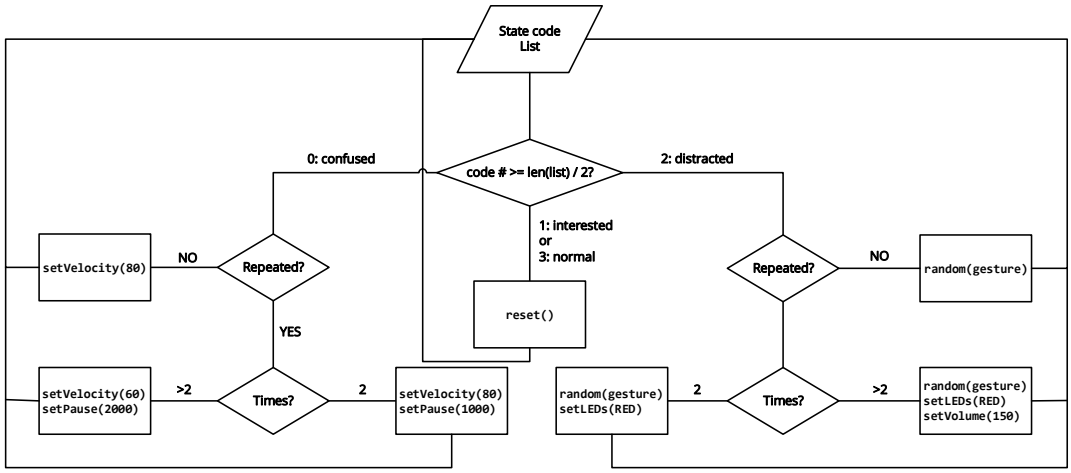



**Figure 26. Mapping strategy**

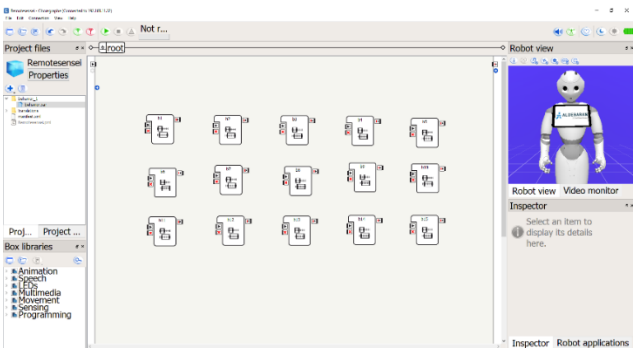
If the "distracted" state is detected, the robot randomly performs one from the desired behaviours. And if this state continues, the robot will flash its LEDs in red. Maintaining this state for the third and more times would trigger an additional increase in its volume.

Also, if the students are confused with the topic, the robot will slow down while speaking. If the state continues, the robot will add longer pauses between sentences.

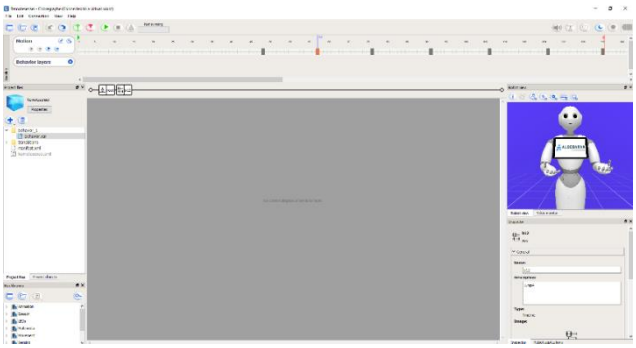
## 3.5 Robotic behaviours generator

### 3.5.1 Design

The design of robotic behaviours was based on data from the Experiment II. These behaviours were designed and simulated in Choregraphe 2.4.3, by the Timeline boxes. The motion of the Pepper robot was composed of posture in each key frame. This research used the technique of Learn from demonstration (LfD) to set parameters of postures by observing human's performance (see Figure 27). The behaviours were summarised in 15 types.




(a)



(b)

**Figure 27. Timeline boxes**

### 3.5.2 Implementation

After the simulation, the designed behaviours were implemented on the Pepper robot. The NAOqi SDK enables the controlling of Pepper's posture, speech, and LEDs. Since the NAOqi SDK only supports Python 2, the code for robot control is developed in Python 2.

Firstly, a proxy is set up for the communication between the computer and the robot. The parameters of designed postures and gestures are exported from Choregraphe.

The name of each joint, with corresponding time and angles, were stored in three lists. Then, the function post.angleInterpolation() takes these lists as the input and controls the robot.

Also, to deliver the lecture, another proxy is created. The robot reads the text through the function say(). The values for volume, speed and pause are defined directly in the passed string.

To light the LEDs on, a new proxy is used. Different colours and durations are set in the function fadeRGB() to realise the flashing of LEDs.

## 3.6 System integration

To establish a connection between the FEA system and the behaviour generator, the project applies publish-subscribe pattern, with the Redis Python library. As the subscriber, the behaviour generator is first started to receive messages from the publisher. It will keep the robot in the normal posture and read the transcript sentence by sentence. However, the loop of listening and the loop of speech cannot run at the same time. In a further improvement, these two loops are assigned to different threads.

At publisher's side, it analyses the current state of the students and looks up the control code

for the state. After acquiring the control code, it will publish it to the behaviour generator.

# Chapter 4: Results and Discussion

## 4.1 Experiment III: System testing

In this section, a series of experiments were conducted on the integrated HRI system. To simulate a teaching scenario, students arranged in groups were asked to sit in front of the robot. The robot, as the lecturer, started with the transcript of a TED talk. To capture students' expressions, the external webcam was connected to the computer and fixed above the robot (see Figure 28). The whole evaluation process went through two phases, which were different in the "action initiator".

**Figure 28. Experimental environment**

### 4.1.1 The first phase: System validation

The first phase was to validate the system. In this case, the students, who intentionally made expressions to trigger the behaviours of the robot, were the "action initiators". The results (see Figure 29) showed that the system was able to correctly recognize the state and control the robot.

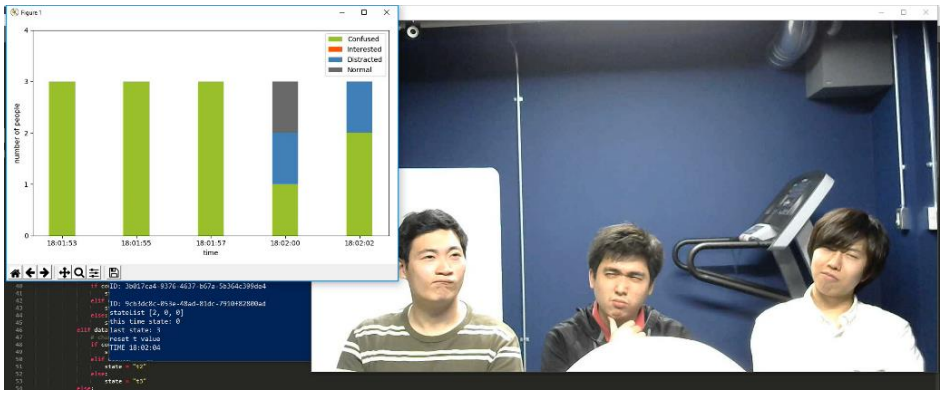



**Figure 29 Students doing "confused" expressions**

### 4.1.2 The second phase: System evaluation

In the second phase, the evaluation process was designed according to the idea of controlled experiment (see Table 4). The students attended the lecture presented by the robot, the "action initiator". Since the experiment was to evaluate the performance of the HRI system, the independent variable was set as the different connection type between the robot and the system.

In the control group, the robot normally delivered a lecture, without interacting with the students. By contrast, the first experimental group connected the robot to the robotic behaviour generator. In this scenario, it performed the desired behaviours randomly, with an interval of 15 seconds. As for the second experimental group, the whole HRI system was involved. While delivering a lecture, the robot performed desired behaviours according to the state detected from the students.

**Table 4: Experimental design**

| | **Control group** | **Experimental group I** | **Experimental group II** |
|---|---|---|---|
| The combination with the HRI system | Without HRI system | With Robotic behaviour generator | With the whole HRI system |

The controlled variable was the content of lecture. A transcript of a 11-minute TED talk was divided into 3 sections and allocated to control group and experimental groups. A total of 16 students participated in the experiment. They were divided into small groups numbered from 1 to 4. For each experiment, the same group of students listened to all 3 sections. Unlike common controlled experiment in biology, the lasting effect in this experiment could be ignored. This project kept the same participants in both control group and experimental group, with the aim of investigating the students' preference for the robot in different groups. In order to reduce the error generating by the attention span, different section was assigned to control group and experimental groups in each experiment (see Table 5). Also, to get desired results, the experiment used transcripts in alternative language for different speakers.

**Table 5: Group allocation**
*(Control group=CG, Experimental group I=EG1, Experimental group II=EG2)*

| **Student group** | **Experimental setting** | | |
|---|---|---|---|
| 1 | CG | EG2 | EG1 |
| 2 | EG1 | CG | EG2 |
| 3 | EG2 | EG1 | CG |
| 4 | CG | EG1 | EG2 |

A questionnaire contained four questions for each section and one for the whole experiment (see Appendix B). After each section, participants were invited to fill out the corresponding part in the questionnaire. The questionnaire was to figure out whether their engagement was improved by the robot. They were asked about the content of the lecture and the behaviour of the robot. As for the questions about the content, correct answers indicated a full understanding of the content.

## 4.2 Experimental Results and Discussion

The data collected by the questionnaire is quantified. The accuracy of answers in each group is weighted by the difficulty level, which is defined by the overall accuracy of questions in each section. The result shows that, to some extent, the HRI system is able to improve the engagement of students.

**Table 6: Results of the questionnaire**

| | **Control group** | **Experimental group I** | **Experimental group II** |
|---|---|---|---|
| Accuracy of answers (%) | 55.5 | 48.3 | 46.3 |
| Degree of engagement (0 to 5) | 0.188 | 2.687 | 3.125 |
| Students who prefer the robot | 2 | 6 | 8 |

As illustrated in Table 6, the accuracy of answers is similar. The accuracy in Control group is 14.9% higher than that in the Experiment group I and 19.8% higher than that in Experiment group II. However, when it comes to the degree of engagement, a significant difference exists between control group and experimental groups. As expected, the group with HRI system shows better performance than the group only with the robotic behaviour generator. Also, when the students are asked about the preference for the robot, most of them chose the robot with the HRI system. Although the accuracy of answers in the control group is slightly higher than the experimental groups, on the whole, the experimental groups excel control group, especially the group with the proposed system.

# Chapter 5: Conclusion and Further Work

To address the potential problem in the current robot-aided education, the project investigates the application of facial emotion recognition and proposes an adaptive HRI system. The main idea is to detect the current state of the students and improve their engagement while they are distracted or confused.

In the development of face expression analysis, the main problems were to track each person in the video stream and to accurately describe the state. To solve the problems, the system combined both expression analysis and face verification to track each person's face. Also, an additional classifier was developed to classify the intensities of emotions into four states. As the method is based on cloud computing and KNN classification, it is possible to run the system without GPUs.

The Pepper robot was controlled by the Robotic Behaviour Generator. The designed behaviours were learned in real educational scenarios. In this project, the relationship between students' state and the lecturer's behaviour was observed, with the assist of the FEA system. Those desired behaviours were collected from the lecturer, during the time period when the students are showing interested. The behaviours are then implemented on the Pepper robot, with Pepper's tool kit. Since the tool kit only support Python 2, the communication problem between FEA system and behaviour generator was solved by socket, which operated in the publish-subscribe pattern.

At the end of the project, a user interface was added. Because of the time caused by facial expression analysis, the execution showed a large delay. Then, to reduce the unexpected delay, the system was improved by separating the UI and the logic of analysis.

To evaluate the system, the project designed a control experiment. In the control group, the robot delivered a lecture without interacting with the students. Although the students' understanding towards the content tends to be better in the control group, on the whole, the HRI system was proved be superior to the systems without interactions.

Also, the proposed HRI system is portable. It can be easily used in the arm robot and the NAO robot, with adjustment of the control code.

The current HRI system is designed for the humanoid robot, which is made up of complex electronic hardware. For further development, an avatar is expected to be added to simply the usage scenario.

# References

Argall, B. D., Chernova, S., Veloso, M. & Browning, B. (2009). A survey of robot learning from demonstration, *Robotics and Autonomous Systems, 57*, no.5, 469-83.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B. & Tanaka, F. (2018). Social robots for education: A review, *Science Robotics, 3*, no.21.

Breazeal, C. (2003). Emotion and sociable humanoid robots, *International Journal of Human-Computer Studies, 59*, no.1-2, 119-55.

Chili (2018, Dec). *CoWriter*. Retrieved 17 April, 2019, from https://chili.epfl.ch/page-92073-en-html/robotics/cowriter/

Elgammal, A., Muang, C. & Hu, D. (2009). Skin Detection - a Short Tutorial, *Encyclopedia of Biometrics, Springer-Verlag Berlin Heidelberg,* 1-10.

Gao, Y. & Leung, M. K. H. (2002). Line segment Hausdorff distance on face matching, *Pattern Recognition, 35*, no.2, 361-71.

Google (2019, Feb). *Vision API - Image Content Analysis | Cloud Vision API | Google Cloud*. Retrieved 5 Feb, 2019, from https://cloud.google.com/vision/

Grekow, J. (2018). *From Content-based Music Emotion Recognition to Emotion Maps of Musical Pieces*, Switzerland, Springer International Publishing.

Lego (2013, Jun). *EV3RSTORM*. Retrieved 17 Apr, 2019, from https://www.lego.com/en-us/mindstorms/build-a-robot/ev3rstorm

Lewis, M., Haviland-Jones, J. M. & Barrett, L. F. (2010). *Handbook of emotions*, Guilford Press.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. In *European conference on computer vision*. 8-16 October, 2016 (Vol.1, pp. 21-37). Amsterdam.

Liu, Z., Wu, M., Cao, W., Chen, L., Xu, J., Zhang, R., Zhou, M. & Mao, J. (2017). A facial expression emotion recognition based human-robot interaction system, *IEEE/CAA Journal of Automatica Sinica, 4*, no.4, 668-76.

Ltutor (2018, May). *What is L2TOR*. Retrieved 6 April, 2019, from http://www.l2tor.eu/what-is-l2tor/

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J. & Ambadar, Z. (2010). The Extended Cohn-

Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression, *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops,* 94-101.

Megvii (2019, March). *Detect API*. Retrieved 30 March, 2019, from https://console.faceplusplus.com.cn/documents/4888373

Microsoft (2019, *Face API - Facial Recognition Software | Microsoft Azure*. Retrieved Feb. 5, from https://azure.microsoft.com/en-us/services/cognitive-services/face/

Ojala, T., Pietikäinen, M. & Harwood, D. (1994). Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *IAPR International Conference on Pattern Recognition*. 9-13 October, 1994 (Vol.1, pp. 582-85). Jerusalem.

Revina, I. M. & Emmanuel, W. R. S. (2018). A Survey on Human Face Expression Recognition Techniques, *Journal of King Saud University - Computer and Information Sciences,*.

Saneiro, M., Santos, O. C., Salmeron-Majadas, S. & Boticario, J. G. (2014). Towards Emotion Detection in Educational Scenarios from Facial Expressions and Body Movements through Multimodal Approaches, *The Scientific World Journal, 2014*, 1-14.

Scikit-Learn (2018, March). *Nearest Neighbors*. Retrieved March 30, 2019, from https://scikit-learn.org/stable/modules/neighbors.html#classification

Softbank (2019, *Pepper*. Retrieved 17 Feb, 2019, from https://www.softbank.jp/en/robot/

Tielman, M., Neerincx, M., Meyer, J. & Looije, R. (2014) Adaptive emotional expression in robot-child interaction. *ACM/IEEE international conference on Human-robot interaction.* Bielefeld.

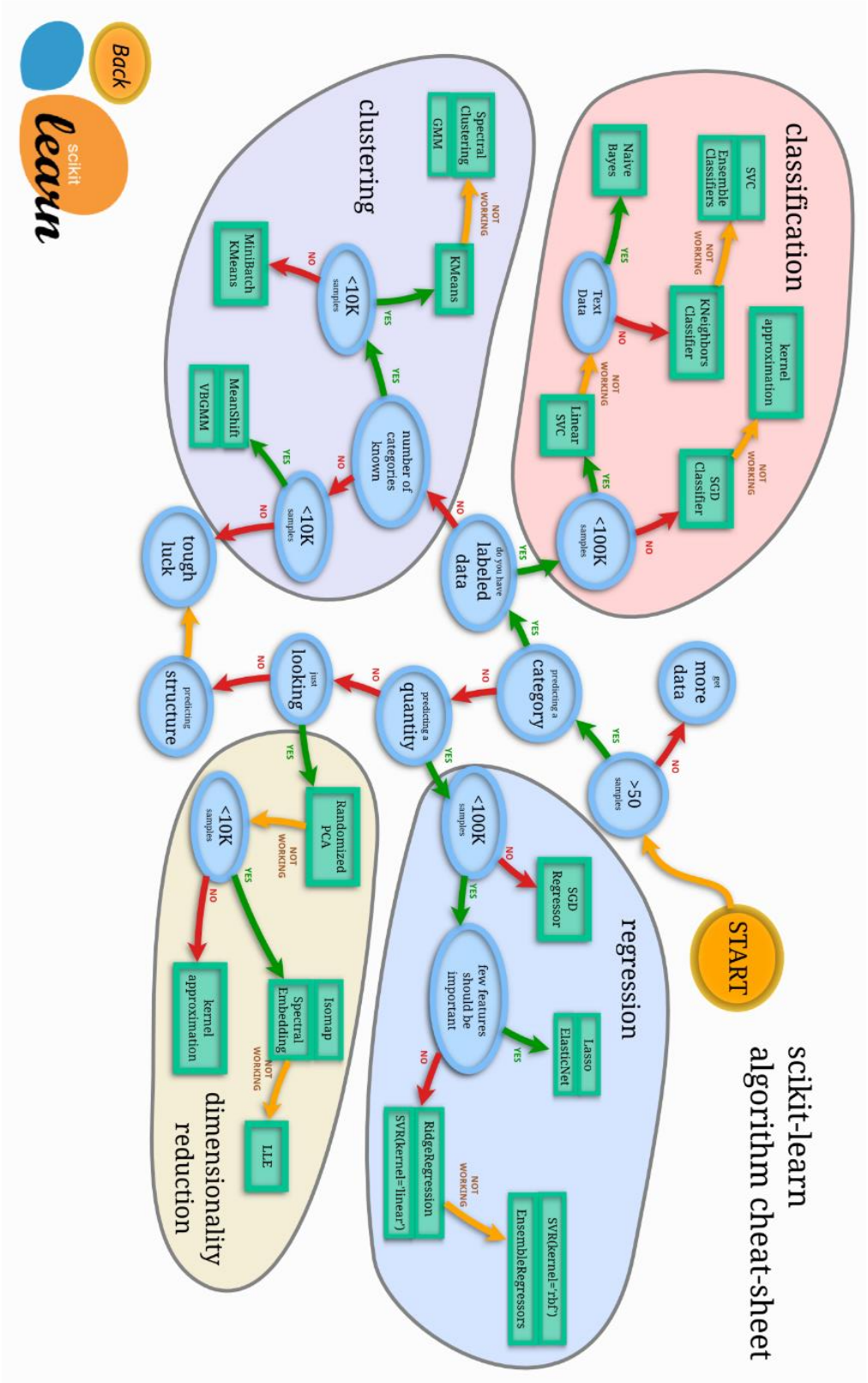
Tracy, J. L. & Randles, D. (2011). Four Models of Basic Emotions: A Review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt, *Emotion Review, 3*, no.4, 397-405.

Viola, P. & Jones, M. (2001) Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition.* Hawaii, IEEE.

Wang, H., Hu, J. & Deng, W. (2018). Face Feature Extraction: A Complete Review, *IEEE Access, 6*, 6001-39.

# Appendix

1. Appendix A: Machine Learning Map



2. Appendix B: Questionnaire (Japanese and English)

# アンケート

1. **学習領域の意味は何ですか?**
   What's the meaning of learning zone?

   - ◯ 人々は彼らが習得したものを再学習する
   - ◯ 人々は習得していないことを学ぶ
   - ◯ 人々は彼らが習得したものに疑問を呈します
   - ◯ 人々は彼らが習得したことを忘れようとします

2. **パフォーマンス領域の意味は何ですか?** *
   What is the meaning of performance zone?

   - ◯ 他人のパフォーマンスから物事を学びます。
   - ◯ 私たちが学んだことを他の人に教える。
   - ◯ できる限り最善を尽くします。
   - ◯ 他の人とパフォーマンスを比較してください。

3. **ロボットの行動の変化に気づきましたか?** *

   - ◯ はい
   - ◯ いいえ(次の質問をスキップ)

4. **変更後、エンゲージメントはどのように改善されましたか? (私は講義にもっと注意を払った)**

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| 全然ない | ◯ | ◯ | ◯ | ◯ | ◯ | たくさん |

5. **デモステネス's 学習領域 で言及されていないものは何ですか？** *
Which thing is not involved in Demosthenes's learning zone?

- ◯ 刀
- ◯ 空気
- ◯ 石
- ◯ 海

6. **どの例が学習領域で言及されていないものは何ですか？** *
Which is not the example of learning zone?

- ◯ セールスマン
- ◯ チェスプレイヤー
- ◯ コンサート
- ◯ 入力

7. **ロボットの行動の変化に気づきましたか？** *

- ◯ いいえ(次の質問をスキップ)
- ◯ はい

8. **変更後、エンゲージメントはどのように改善されましたか？ （私は講義にもっと注意を払った）**

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| 全然ない | ◯ | ◯ | ◯ | ◯ | ◯ | たくさん |

9. **講師によると、彼はどの領域でもっと時間を過ごすべきだと思いますか？** *

- ◯ 学習領域
- ◯ パフォーマンス領域

10. **講師によると、「学校」はどの領域にあるべきですか？** *

- ◯ 学習領域
- ◯ パフォーマンス領域

11. **ロボットの行動の変化に気づきましたか?** *

- ◯ はい
- ◯ いいえ(次の質問をスキップ)

12. **変更後、エンゲージメントはどのように改善されましたか? (私は講義にもっと注意を払った)**

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| 全然ない | ◯ | ◯ | ◯ | ◯ | ◯ | たくさん |

13. **どのロボットが好きですか?** *

- ◯ セクション1のロボット
- ◯ セクション2のロボット
- ◯ セクション3のロボット

## Questionnaire-2

Please fill in the blank according to your own understanding.

1. **What's the meaning of learning zone?**

- ◯ The zone people relearn things they have mastered
- ◯ The zone people learn things that they haven't mastered
- ◯ The zone people question what they have mastered
- ◯ The zone people try to forget what they have mastered

2. **What is the meaning of performance zone?** *

- ◯ Learn things from others' performance.
- ◯ Teach others what we have learned.
- ◯ Do something as best as we can.
- ◯ Compare the performance with others.

3. **Did you notice the change of the robot's behaviour?** *

- ◯ Yes
- ◯ No (skip next question)

4. **How did your engagement improve after the change ? (I paid more attention to the lecture)**

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Not at all | ◯ | ◯ | ◯ | ◯ | ◯ | A lot |

5. **Which thing is not involved in Demosthenes's learning zone?** *

- ◯ sword
- ◯ air
- ◯ stone
- ◯ sea

6. **Which is not the example of learning zone?** *

- ◯ salesman
- ◯ chess player
- ◯ concert
- ◯ typing

7. **Did you notice the change of the robot's behaviour?** *

- ◯ No (skip next question)
- ◯ Yes

8. **How did your engagement improve after the change ? (I paid more attention to the lecture)**

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Not at all | ◯ | ◯ | ◯ | ◯ | ◯ | A lot |

9. **According to the lecturer, which zone does he think we should spend more time in?** *

- ◯ learning zone
- ◯ performance zone

10. **According to the lecturer, the "school" is in which zone?** *

- ◯ learning zone
- ◯ performance zone

11. **Did you notice the change of the robot's behaviour?** *

- ◯ No (skip next question)
- ◯ Yes

12. **How did your engagement improve after the change ? (I paid more attention to the lecture)**

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Not at all | ◯ | ◯ | ◯ | ◯ | ◯ | A lot |

13. **Which robot do you prefer?** *

- ◯ Robot in Section A
- ◯ Robot in Section B
- ◯ Robot in Section C

# 北京邮电大学 本科毕业设计（论文）任务书

# Project Specification Form

## Part 1 - Supervisor

| | | | |
|---|---|---|---|
| 论文题目<br>Project Title | Facial emotion recognition applied to a humanoid robot teacher | | |
| 题目分类<br>Scope | Multimedia and Vision | Research | Software |
| 主要内容<br>Project description | The project aims to study facial emotion recognition technology, and find out how it can be applied to a humanoid robot teacher. As technology to recognise faces and to detect basic emotions from facial expressions is steadily improving, having a robot that observes the students and can notify the lecturer or modify its own displayed emotions when changes in the students' emotions are detected, would be an advantage. The project is thus about programming robots (e.g. NAO, Pepper) to react to the students' emotions in real time in order to improve students' engagement and sustain learning. The robot should be able to adapt its own emotions (conveyed through social signals such as head poses, body postures and tone of voice) in response to the detected emotions in the audience. If time allows, two versions may be implemented, one using a real robot and one using a virtual version of the robot (avatar) in an online lecture. | | |
| 关键词<br>Keywords | Emotion recognition, Teacher robots, Robots emotional expressions, Avatars | | |
| 主要任务<br>Main tasks | 1 Research about facial emotion recognition technology | | |
| | 2 Data collection and emotion recognition system implementation | | |
| | 3 Robot's emotional expressions implementation | | |
| | 4 System evaluation | | |
| 主要成果<br>Measurable outcomes | 1 Facial emotion recognition system | | |
| | 2 System that can relate students' emotions with the robot's emotions | | |
| | 3 Adaptive emotional teaching robot | | |

# 北京邮电大学 本科毕业设计（论文）任务书

# Project Specification Form

# Part 2 - Student

| | | | | | |
|---|---|---|---|---|---|
| **学院 School** | International School | **专业 Programme** | **Telecommunications Engineering with Management** | | |
| **姓 Family name** | Shi | **名 First Name** | Yuyuan | | |
| **BUPT 学号 BUPT number** | 2015212967 | **QM 学号 QM number** | 151009631 | **班级 Class** | 2015215104 |
| **论文题目 Project Title** | Facial emotion recognition applied to a humanoid robot teacher | | | | |
| **论文概述 Project outline**<br><br>**Write about 500-800 words**<br><br>**Please refer to Project Student Handbook section 3.2** | **Objectives**<br><br>This project aims to study facial emotion recognition and to find out how it can be applied to a humanoid robot teacher in order to improve students' engagement and sustain learning.<br><br>The expected humanoid robot teacher could not only deliver lectures but also adapt its teaching style to the emotion of students. Their emotion is judged by the facial expressions. When negative emotions (such as anger, contempt, sadness) occur, the robot teacher is expected to take actions that could help with students' learning efficiency. These actions include changing head poses, body postures and tone of voice which human beings will do to express their emotion.<br><br>The whole system consists of two major parts, one is the real-time emotional recognition module, the other is the robot's feedback module. Also, an effective mechanism is required to relate the emotion of the students to that of the robot teacher, so that we are able to receive instructions from the feedback module as a proper response.<br><br>**Methodology**<br><br>As for the emotional recognition module, the current technology is advanced enough for its implementation. The applicable tool, for example, Face API from Microsoft s cognitive service could be used for the detection for basic emotions. Combining with the package opencv-python, we could get frames from the camera and request the Face API. After analysing, one's expression would be scored in several dimensions, such as: happiness, sadness, surprise, anger, fear, contempt, disgust, and neutral. Since some of these emotions, like disgust and fear, may not be suitable to describe students' emotion in class, a more specific model is needed.<br><br>To research about how the behaviour of the teacher could influence the students' emotions, it's possible for us to collect students' face data in class, using the previous recognition module to analyse their emotion. We manage to play lecture videos (e.g. TED talk) to students and observe their reactions. And | | | | |

finally, mark the points where students show positive emotion and record the corresponding behaviour (expression, tone and gesture) of the teacher.

In order to improve the students' engagement and sustain learning, we are supposed to study about the desired behaviour that a robot should have. It would change its teaching style to attract students' attention. And its tone or gestures could be more exaggerated than usual. Or more directly, it could just provide students with a reminder. The behaviours of the lecturer that we have picked from videos could assist us with the design of the robot's actions. In addition, questionnaires are also helpful to know about students' thoughts.

**Experiments**

The first experiments would be carried out to verify the usability of the emotional recognition module. We are going to make some expressions on purpose to test the analytical accuracy of the module.

The second experiment aims at relating students' emotion to the behaviour of the lecturer. We will use the emotional recognition system to analyse the emotion of the students when watching lectures, and produce a graphical output. The output is a trend chart showing the variation of students' emotions in response to lecturers' behaviours. Focus on time when their emotion varies from negative to positive and vice versa, we are able to get a set of expected behaviours for the robot teacher.

The final experiment is to evaluate the whole system. We will do a scenario testing to verify the performance of the whole system. If possible, we can put the robot into a real class, check the students emotion analysed by the system and then the corresponding actions taken by the robot.

**Software / Hardware**

The whole system is going to be developed in Python. Microsoft Face API would be used in the implementation of emotional recognition system. Some packages or module like opencv-python, NumPy, Matplotlib and so forth would also be imported. In addition, Python SDK for NAOqi is needed.

The hardware included is the Pepper robot.

**A list of background material consulted including World Wide Web pages**

[1] Zhentao, L., Min, W., Weihua, C., Luefeng, C., Jianping, X., Ri, Z., Mengtian, Z., and Junwei, Mao. (2017). A Facial Expression Emotion Recognition Based Human-robot Interaction System. IEEE/CAA Journal of Automatica Sinica. vol. 4, no. 4.

[2] Mohamed S., Sofia G J. (2013). Effect of facial expressions on student's comprehension recognition in virtual educational environments. SpringerPlus. 2:455

[3] Ron S., (2017). *Does Emotive Computing Belong in the Classroom?* [online] Available from: https://www.edsurge.com/news/2017-01-04-does-emotive-computing-belong-in-the-classroom [Accessed 26 Oct 2018]

<table>
<tr><td></td><td>[4] Melissa Jun R., (2017). The rise of robot teachers [online] Available from: https://newsroom.cisco.com/feature-content?articleId=1873531 [Accessed 28 Oct 2018]<br><br>[5] Microsoft, (2017). How to Analyze Videos in Real-time [online] Available from: https://docs.microsoft.com/en-us/azure/cognitive-services/emotion/emotion-api-how-to-topics/howtoanalyzevideo_emotion [Accessed 2 November 2018]<br><br>[6] Microsoft, (2017). Face API Documentations [online] Available from: https://azure.microsoft.com/zh-cn/services/cognitive-services/ [Accessed 2 November 2018]<br><br>[7] Aldebaran, (2018) NAOqi - Developer guide [online] Available from: http://doc.aldebaran.com/2-4/index_dev_guide.html [Accessed 6 November 2018]</td></tr>
<tr><td rowspan="2">道德规范<br>Ethics</td><td>Please confirm that you have discussed ethical issues with your Supervisor using the ethics checklist on QMPlus.<br>[YES/<s>NO</s>]</td></tr>
<tr><td>Summary of ethical issues: (put N/A if not applicable)<br><br>1. Will the participants be exposed to any risks greater than those encountered in their normal working life?<br>No. The participants won’t be exposed to any risks. They only have to sit down and attend a lecture as they usually do. They will be facing a robot but there will be no direct contact with it.<br><br>2. Will the participants be using any non-standard hardware?<br>Yes. A robot will interact with participants.<br><br>3. How will participants voluntarily give consent?<br>They will give a verbal agreement.<br><br>4. Are you offering any incentive to the participants?<br>No.<br><br>5. Is there any intentional deception of the participants?<br>No.<br><br>6. Are any of your participants under the age of 16?<br>No.<br><br>7. Do any of your participants have an impairment that will limit their understanding or communication?</td></tr>
</table>

| | |
|---|---|
| | No.<br><br>8. Are you in a position of authority or influence over any of your participants?<br>No. They are all college students.<br><br>9. Will the participants be informed that they could withdraw at any time?<br>Yes.<br><br>10. Will the participants be informed of your contact details?<br>Yes.<br><br>11. Will the participants be debriefed?<br>Yes. While recruiting participants, we will explain the objective of our project and how their expression information will be used to them.<br><br>12. Will the data collected from participants be stored in an anonymous form?<br>Yes. |
| **中期目标**<br>**Mid-term target.**<br><br>**It must be tangible outcomes, E.g. software, hardware or simulation.**<br><br>**It will be assessed at the mid-term oral.** | Finish the facial recognition system and evaluate it.<br>Use the result of experiment to refine the system.<br>Complete the dataset of desired behaviours of the robot teacher. |

# Work Plan (Gantt Chart)

Fill in the sub-tasks and insert a letter X in the cells to show the extent of each task

| | Nov | Dec | Jan | Feb | Mar | Apr | May |
|---|---|---|---|---|---|---|---|
| **Task 1 Research about facial emotion recognition technology** | | | | | | | |
| Background reading | X | X | | | | | |
| Research about available expression analysis tools | X | X | | | | | |
| | | | | | | | |
| | | | | | | | |
| **Task 2 Data collection and emotion recognition system implementation** | | | | | | | |
| Data collection | X | X | X | | | | |
| Emotion recognition system implementation | | X | X | | | | |
| Testing | | | X | | | | |
| Refining | | | X | | | | |
| **Task 3 Robot's emotional expressions implementation** | | | | | | | |
| Analyse the face data of students | X | X | X | | | | |
| Collect positive behaviours of the teacher | | X | X | | | | |
| Design the behaviours of the robot teacher | | | | X | X | X | |
| | | | | | | | |
| **Task 4 System evaluation** | | | | | | | |
| Plan for the evaluation demonstration | | | | | X | | |
| Contact participants | | | | | X | X | |
| System evaluation | | | | | | X | X |
| | | | | | | | |

# 北京邮电大学 本科毕业设计（论文）初期进度报告

# Project Early-term Progress Report

| | | | | | |
|---|---|---|---|---|---|
| **学院 School** | International School | **专业 Programme** | **Telecommunications Engineering with Management** | | |
| **姓 Family name** | Shi | **名 First Name** | Yuyuan | | |
| **BUPT 学号 BUPT number** | 2015212967 | **QM 学号 QM number** | 151009631 | **班级 Class** | 2015215104 |
| **论文题目 Project Title** | Facial emotion recognition applied to a humanoid robot teacher | | | | |

**已完成工作 Finished work:**

**Materials read or researched**

1. Facial emotion recognition technology in human-robot interaction
   Facial expression emotion recognition (FEER) has been studied for a long time. Many researchers combine this technology with human-robot interaction system, in order that the robot could provide natural and appropriate response. For example, some researchers use FEER to make robot generate corresponding facial expression to smooth the conversation [1], while some apply it in analysing underlying semantic structures and identify topics in conversations [2]. Building on these ideas, we could further equip the robot with a more adequate emotion expression system.

2. Affective artificial intelligence in education (AIED)
   To make the system more practical, I also referred to existing research about the relationship between emotion and learning, which was more related to psychology field. According to a paper that reviews and integrates researches related to the AIED system [3], there exists complex interplay between affect and learning-related cognitive processes.

   The emotion of human, to some extent, affecting judgement, attention, motivation, problem solving and so forth. Students who are in a positive emotion before working on a task would view the task as more interesting than individuals in a negative or neutral mood. This should lead to better performance and deeper satisfaction along with greater effort expenditure. However, negative emotions, such as boredom and sadness, would negatively influence learning efficiency.

3. Existing facial expression technologies
   I read through the documents of several popular emotion analysis tools and compared them at the initial stage.
   3.1. Microsoft Face API
   - Analysis results in 8 dimensions: happiness, sadness, surprise, anger, fear, contempt, disgust, and neutral.
   - Each dimension scores from 0 to 1.

   3.2. Google Vision API
   - Analysis results in 4 dimensions: joy, sorrow, anger, surprise.
   - Each dimension is evaluated in 5 degrees (except ‘Unknown’): Very Unlikely, Unlikely, Possible, Likely, Very Likely.

   3.3. OpenPose model
   - Provide key points of each face. The model requires further training to recognize emotion from expressions.

- The emotion types are custom and a large dataset is needed.

Considering the accuracy and practicality, I decided to choose Microsoft Face API, since Google's API is not suitable for quantifying the emotion and the use of OpenPose model requires a large size of dataset (facial information at class).

4. Related materials of the robot
   I read through the document of Pepper robot [4] and identified the postures that the Pepper could perform. It can be seen from the document that Pepper's shoulder yaw, shoulder roll, elbow yaw and elbow roll enable him to move with various postures.

- **Completed work**

1. Face data collection
   The dataset covers videos recorded in lectures and simulated scenes. Additional data may need for further refinement.

2. Implementation of expression analysis system
   The system has been improved to be able to track each person's emotion variation.
   To realize this function, I divide whole process into 3 steps: face detection, face comparing and data recording. The *Face-Detect* API and *Face-Find Similar* API are used. The flow chart is shown as below:

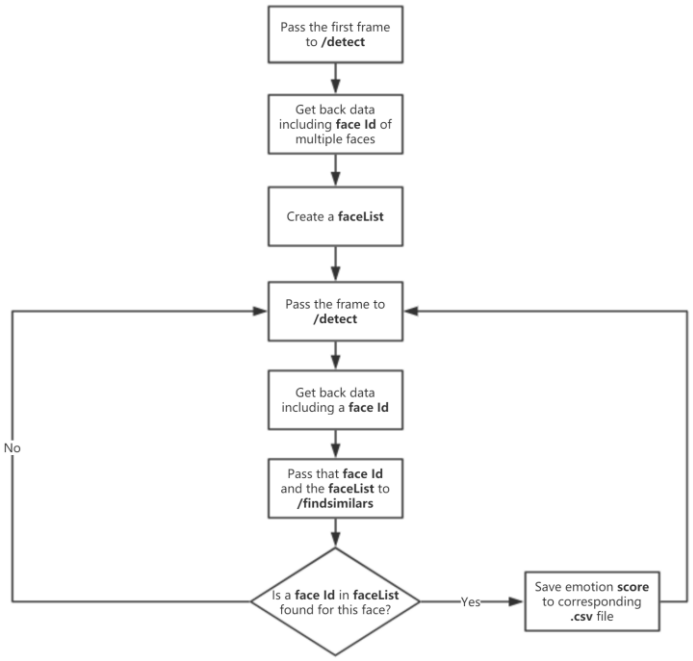



*Figure 30 Flow Chart*

Analysis results are stored in csv files:

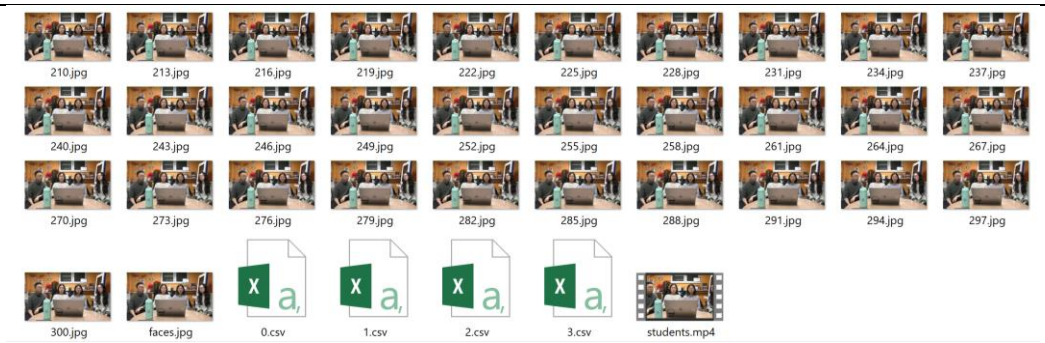



*Figure 31 Frames and csv files*

Transform the data to line graphs:

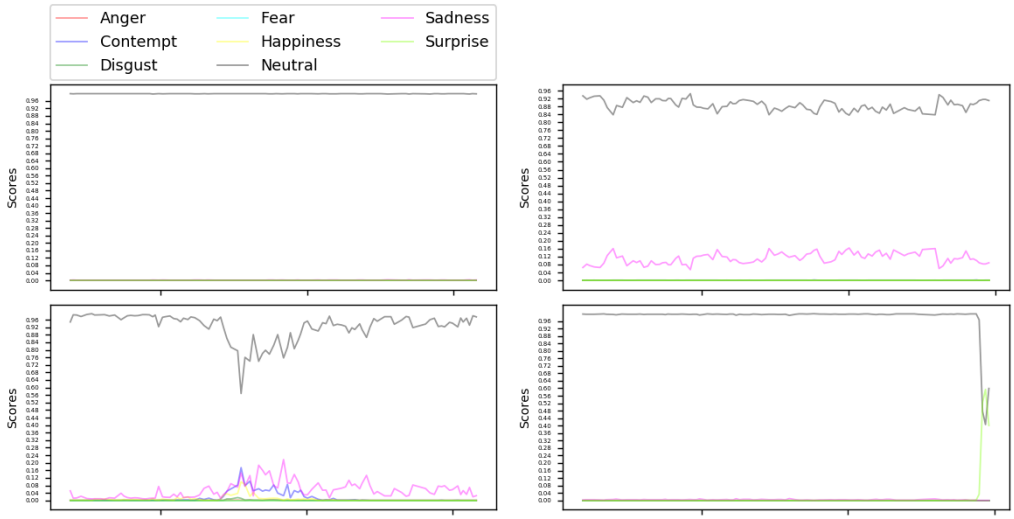



*Figure 32 Line graphs*

3. The first experiment—verifying the usability of the emotional recognition module
Students were asked to perform several expressions on purpose: when they were thinking, confused, distracted and showing interest.

| Expression | Picture | Result |
|---|---|---|
| **Thinking** |  | {'anger': 0.0, 'contempt': 0.001, 'disgust': 0.0, 'fear': 0.0, 'happiness': 0.003, 'neutral': 0.961, 'sadness': 0.035, 'surprise': 0.0}<br>{'anger': 0.0, 'contempt': 0.0, 'disgust': 0.0, 'fear': 0.0, 'happiness': 0.004, 'neutral': 0.995, 'sadness': 0.001, 'surprise': 0.0}<br>{'anger': 0.0, 'contempt': 0.0, 'disgust': 0.0, 'fear': 0.0, 'happiness': 0.0, 'neutral': 0.861, 'sadness': 0.139, 'surprise': 0.0} |
| **Confused** |  | {'anger': 0.0, 'contempt': 0.004, 'disgust': 0.0, 'fear': 0.0, 'happiness': 0.0, 'neutral': 0.979, 'sadness': 0.017, 'surprise': 0.0}<br>{'anger': 0.0, 'contempt': 0.032, 'disgust': 0.002, 'fear': 0.0, 'happiness': 0.031, 'neutral': 0.885, 'sadness': 0.05, 'surprise': 0.0}<br>{'anger': 0.0, 'contempt': 0.0, 'disgust': 0.0, 'fear': 0.0, 'happiness': |

| | | |
|---|---|---|
| | | 0.0, 'neutral': 0.958, 'sadness': 0.042, 'surprise': 0.0} |
| **Distracted** |  | {'anger': 0.0, 'contempt': 0.002, 'disgust': 0.0, 'fear': 0.0, 'happiness': 0.001, 'neutral': 0.988, 'sadness': 0.009, 'surprise': 0.0}<br>{'anger': 0.0, 'contempt': 0.003, 'disgust': 0.0, 'fear': 0.0, 'happiness': 0.001, 'neutral': 0.956, 'sadness': 0.04, 'surprise': 0.0}<br>{'anger': 0.0, 'contempt': 0.0, 'disgust': 0.0, 'fear': 0.0, 'happiness': 0.0, 'neutral': 0.998, 'sadness': 0.002, 'surprise': 0.0} |
| **Showing interest** |  | {'anger': 0.0, 'contempt': 0.0, 'disgust': 0.0, 'fear': 0.0, 'happiness': 1.0, 'neutral': 0.0, 'sadness': 0.0, 'surprise': 0.0}<br>{'anger': 0.0, 'contempt': 0.0, 'disgust': 0.0, 'fear': 0.0, 'happiness': 1.0, 'neutral': 0.0, 'sadness': 0.0, 'surprise': 0.0}<br>{'anger': 0.0, 'contempt': 0.0, 'disgust': 0.0, 'fear': 0.0, 'happiness': 0.997, 'neutral': 0.003, 'sadness': 0.0, 'surprise': 0.0} |

From the analysis results, we could notice that thinking behaviour seems complex when related to expression, since students show either positive or negative expression while thinking. However, analysis of other behaviours has shown desired result, for example, when students get confused, the total score of negative emotions is larger than that of positive emotions.

4. The second experiment— relating students' emotion to the behaviour of the lecturer
   By marking the video frames, some desired behaviours were collected. More work needed to be done to complete the whole set.

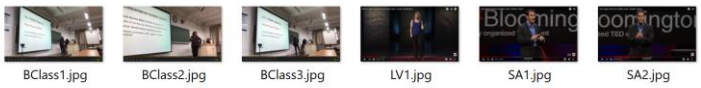



*Figure 33 Marked behaviours*

- **Problems**

1. Every time when I passed the same face to *Face-Detect* API, I got a different face Id, so that I was not able to track each face.

2. The shooting angle was not desired so that I got unwanted analysis results.

**Solutions**

1. I applied *Face-Find Similar* API to match a new face Id to the existing face list, so that I was able to recognize faces of the same person and tracked each face.

2. I changed the position of camera in the later data collection, setting it right in front of the students.

[1] Zhentao, L., Min, W., Weihua, C., Luefeng, C., Jianping, X., Ri, Z., Mengtian, Z., and Junwei, Mao. (2017). "A Facial Expression Emotion Recognition Based Human-robot Interaction System," in *IEEE/CAA Journal of Automatica Sinica*. Vol. 4, pp. 668-676
[2] Li, Z., Ming, J., Dewan F., M.A. H. (2013) "Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot," in *Expert Systems with Applications*. Vol. 40, pp. 5160-5168.
[3] Emmanuel G. B., Boris V., Yuan-Jin H., Susanne P. L. (2009) "Affective Artificial Intelligence in Education: From Detection to Adaptation," in *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling.* Pp. 81-88
[4] Aldebaran, (2018) NAOqi - Developer guide [online] Available from: http://doc.aldebaran.com/2-4/family/pepper_technical/index_dev_pepper.html [Accessed 30 December 2018]

**是否符合进度? On schedule as per GANTT chart?**
[YES/NO] YES

**下一步 Next steps:**

1. Keep on carrying out the experiment of relating students' emotion to the behaviour of the lecturer.
2. Complete the set of desired behaviours.

# 北京邮电大学 本科毕业设计（论文）中期进度报告

# Project Mid-term Progress Report

| | | | | | |
|---|---|---|---|---|---|
| **学院 School** | International School | **专业 Programme** | **Telecommunications Engineering with Management** | | |
| **姓 Family name** | Shi | **名 First Name** | Yuyuan | | |
| **BUPT 学号 BUPT number** | 2015212967 | **QM 学号 QM number** | 151009631 | **班级 Class** | 2015215104 |
| **论文题目 Project Title** | Facial emotion recognition applied to a humanoid robot teacher | | | | |

**是否完成任务书中所定的中期目标？Targets met (as set in the Specification)?**
[YES/NO] YES

**已完成工作 Finished work:**

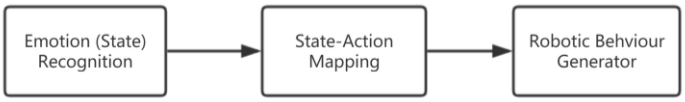



*Figure 34 System structure*

## 1. Data collection (Task 2-1)

The database consists of videos and labelled faces. The videos are collected in two scenarios: real teaching and simulated scenes. In the real teaching scenarios, both teacher and students' behaviours are recorded. As for the simulation, the speech videos are played to a group of (3 to 5) students and their reactions are recorded.

The labelled faces are also from two sources, one is the posed expressions of how we usually act during the lecture, the other is the extended Cohn-Kanade Dataset (CK+) [1, 2]. The CK+ dataset is composed by, for each emotion, the sequences from the neutral face to the peak expression.

## 2. Facial expression analysis (Task 2-2)

In this research, facial expression analysis (FEA) is the core part of emotion recognition system. It is tested in the Experiment I, by applying it to the labelled faces. And it is also used in the Experiment II, which relates the students' emotional state to the teacher's behaviour. So far, the system is able to track each person's emotion in a video or webcam view.

Microsoft Face API provides face verification, face detection and emotion recognition for both single and group photos. After requesting the API, it would return attributes of each face, including landmarks and emotion. The analysis results in 7 basic emotions (Happiness, Surprise, Disgust, Sadness, Anger, Fear, and Contempt) and Neutral. Each emotion is scored from 0 to 1.

Face detection is realised by *Face-Detect* API. Since it only accepts image data, more operations are required for the video analysis. First, OpenCV-Python package is used to extract frames from the video file or the camera. This operation is performed by

cv2.VideoCapture() and read() method. Then, we would get a series of analysis results for faces in each frame. However, according to the mechanism of the *Face-Detect* API, each time when the frame is passed to the API, faces will be marked with unique face IDs. Because different face IDs are assigned to faces of the same person, tracking faces within the video seems impractical. Fortunately, the *Face-Detect* API provides a possible solution.

Face verification is carried out by *Face-Find Similar* API. By requesting the API with the query face ID and an array of candidate face IDs, we could get a matched candidate face ID with a confidence value. As in the educational scenario, the students are sitting without exaggerated movement, so that the accuracy of matching is guaranteed.

The tracking process then results in three phases: creating a face list, matching the face and recording analysis results for each person. At the beginning, the first frame is passed to the *Face-Detect* API. The returned face IDs are then saved as the array of candidates. For the following frames, each face within will be matched to a specific face ID. The analysis results of the same face ID, namely the same person, would be stored into a specific csv file. Also, the timestamp for each frame is listed in the files for the further locating.

## 3. Experiment I: Verifying the usability of the emotional recognition module (Task2-3,4)

To verify the accuracy of the FEA tool, we apply it to the CK+ database. The code is modified for carrying out a batch processing. Within each batch, the faces range from neutral to the peak expression. It could be seen from the result that, the more exaggerated expression, the better accuracy. Take the first batch as example (Figure 2), the turning point is at around the fifth level. And the analysis result (Table 1) is exactly matched with the provided label (Figure 3).

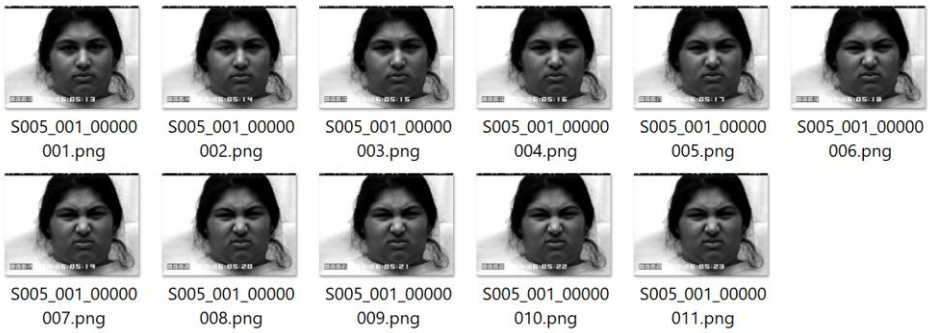



*Figure 35 The first sequence of faces*

| | anger | contempt | disgust | fear | happiness | neutral | sadness | surprise |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.004 | 0.003 | 0 | 0 | 0 | 0.962 | 0.03 | 0 |
| 1 | 0.006 | 0.005 | 0 | 0 | 0 | 0.965 | 0.024 | 0 |
| 2 | 0.014 | 0.011 | 0.001 | 0 | 0 | 0.955 | 0.018 | 0 |
| 3 | 0.037 | 0.01 | 0.002 | 0 | 0 | 0.947 | 0.004 | 0 |
| 4 | 0.345 | 0.004 | 0.512 | 0 | 0 | 0.139 | 0.001 | 0 |
| 5 | 0.28 | 0.001 | 0.716 | 0 | 0 | 0.004 | 0 | 0 |
| 6 | 0.271 | 0 | 0.728 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0.107 | 0 | 0.893 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0.129 | 0 | 0.87 | 0 | 0 | 0.001 | 0 | 0 |
| 9 | 0.129 | 0 | 0.871 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0.084 | 0 | 0.915 | 0 | 0 | 0 | 0 | 0 |

*Table 7 Analysis results for data in S005\001*

```
------------------------------------------------------------
Emotion code at:

Emotion/S005/001/S005_001_00000011_emotion.txt

which has
3.00000000e+00

that is disgust
------------------------------------------------------------
```

*Figure 36 Description in the readme file*

The module is also tested in real-time with the webcam. However, the results showed that the Neutral emotion often got high value even when we performed some basic emotions (e.g. sad, happy). From the earlier research, we know that if the expression is not obvious, it is likely to be identified as Neutral. To make a further improvement, a higher level of emotions is taken into consideration.

Instead of basic ones, the higher level of emotions is more of describing students' current state. Here, we use the other labelled dataset. To construct this dataset, students perform in the way when they are thinking, confused, distracted and showing interest. Faces of different states are analysed by the emotional recognition module, and then labelled with different emotion scores. The emotion scores are considered as the features of those labelled faces.

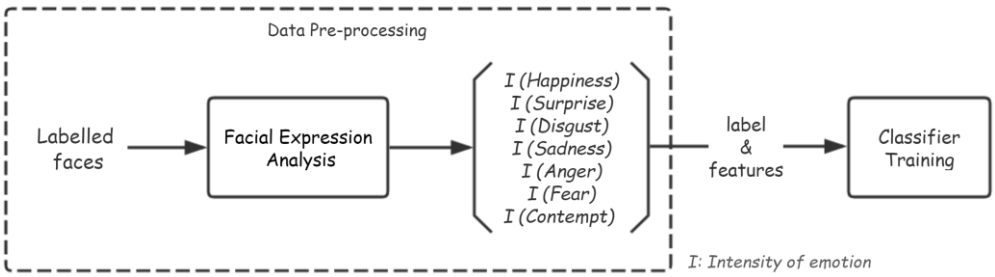



*Figure 37 Experiment I process*

Then, a classifier is developed to take the scores of basic emotions (features) as input, and output the specific state (label). To begin with, we use the K-Nearest Neighbours (KNN) algorism to train our classifier. For each kind of states, we have 20 pictures for training and 10 for testing.

## 4. Experiment II: Relating students' emotion to the behaviour of the lecturer (Task 3-1,2)

In this section, we will use the video dataset. While analysing the dataset, we use the cv2.CAP_PROP_POS_MSEC attribute to mark the current position of the video file. The resulting csv file for each person contains the timestamp and scores of basic emotions. During the interval where desired state (i.e. thinking, confused and showing interest) occurs on more than half of students, the teacher's behaviours are recorded, by the description of gesture, head pose and tone of voice.

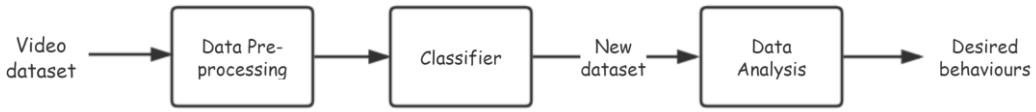



*Figure 38 Experiment II process*

## 5. Designing the robotic behaviours

The design work is realised by reference to existing emotion expression models and the data collected from Experiment II. Before the final implementation, the design is represented through simulation tool embedded in the *Choregraphe* software.
To start with, we simply consider positive robotic behaviours as the objective of increasing students' efficiency in study. The nonverbal behaviours are based on the work of Häring et al. [3] and Beck et al. [4] And the verbal behaviours are based on Mythe's model [5], which relates the arousal of the emotion to the fundamental frequency of voice, speech rate and speech volume of the voice.
Besides human-designed work, those behaviours learned from human are also turned into parameters (Figure 6).

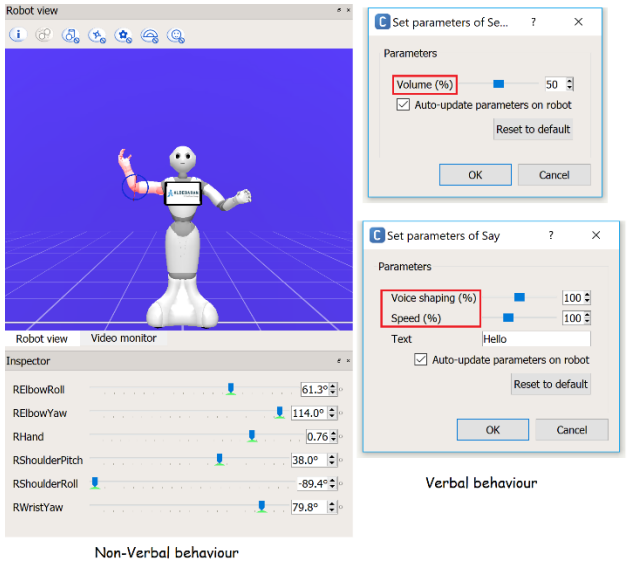



*Figure 39 Setting of parameters*

[1] Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), Grenoble, France, 46-53.
[2] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression. Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010), San Francisco, USA, 94-101.
[3] Häring M, Bee N, André E (2011) Creation and Evaluation of emotion expression with body movement, sound and eye color for humanoid robots. In: The 20th IEEE international symposium on robot and human interactive communication, RO-MAN, pp 204– 209. https://doi.org/10.1109/ROMAN.2011.6005263
[4] Beck A, Canamero L, Bard KA (2010) Towards an affect space for robots to display emotional body language. In: Proceedings of the 19th IEEE international symposium on robot and human

interactive communication, RO-MAN. IEEE Press, Piscataway, New Jersey, pp 464–469. https://doi.org/10.1109/ROMAN.2010.5598649
[5] Tielman, M., Neerincx, M., Meyer, J. J., & Looije, R. (2014). Adaptive emotional expression in robot-child interaction. In Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction, pp. 407-414.

**尚需完成的任务 Work to do:**

## 1. System Integration

a) Implement the designed behaviours on the Pepper robot
To reduce the time of response, the designed robotic behaviours are turned into Python code and then loaded to Choregraphe.
b) Connect different modules
Because of the limitation of hardware, more work is needed for coordinating the robot, program and webcam.

## 2. Evaluate the whole system.

a) Prepare
b) Demonstrate
c) Analyse results

**存在问题 Problems:**

**1. Behaviour data size is too small.**
We only obtained a small number of desired behaviours from *Experiment II: Relating students' emotion to the behaviour of the lecturer*. The judgment of *good state* in this experiment is used to base on the occurrence of positive *basic emotions* (e.g. happiness, surprise) before the refinement. After changing the criteria to the positive *states* (e.g. thinking, showing interest), the size of data becomes much smaller.

**2. The designed behaviours are limited**
We only considered the behaviours that could be performed by humans.

**拟采取的办法 Solutions:**

1. Try to enhance the accuracy of classifier by adopting other training models. Or try to figure out a better description of overall status based on individual status. If possible, we could also attempt to get more video data.
2. Consider the behaviours that are defined as robot-specific nonverbal behaviour (RNB), for example, the LED colour.

**论文结构 Structure of the final report:**

**Abstract**
The brief summary of the research's purpose, approach, results and conclusions.

**Chapter 1: Introduction**
An introductory part of the research, which contains:
- Related work
- Research gap and its importance

- Overview of the methodology
- Structure of the paper
- Main conclusions

**Chapter 2: Background**

1. Overview of current robot-aided education
2. State-of-art facial expression recognition
   a. General recognition process
   b. Existing analysis tools
3. Robotic emotion expression
   a. Emotion state models and frameworks
   b. Behaviours learned from human
4. The Pepper robot

**Chapter 3: Design and Implementation**

1. Construction of the video dataset
   a. Collection
   b. Annotation
2. Implementation of facial expression recognition system
3. Experiment I: Applying the emotion recognition system
4. Experiment II: Relating the human emotion to the teacher's behaviours (robot learning from demonstration)
5. Design of the emotion expression system (behaviours of the robot)
   a. Mapping strategy (state-action)
   b. Implement the design on the Pepper robot
6. Experiment III: Evaluating the system in the educational scenario

**Chapter 4: Results and Discussions**

1. Results of Experiment I: discuss the accuracy of the recognition system
2. Results of Experiment II: discuss the mapping relationship between student's emotional states and teacher's behaviours
3. Evaluation results of the whole system (Experiment III)
4. Comparing the results with desired outcomes

**Chapter 5: Conclusion and Further Work**

1. Conclusion: summary of the results
2. Further work: possible improvement

# 北京邮电大学 本科毕业设计（论文）教师指导记录表

# Project Supervision Log

| | | | | | |
|---|---|---|---|---|---|
| **学院 School** | International School | **专业 Programme** | **Telecommunications Engineering with Management** | | |
| **姓 Family name** | Shi | **名 First Name** | Yuyuan | | |
| **BUPT 学号 BUPT number** | 2015212967 | **QM 学号 QM number** | 151009631 | **班级 Class** | 2015215104 |
| **论文题目 Project Title** | Facial emotion recognition applied to a humanoid robot teacher | | | | |

Date: 19-10-2018
Supervision type: face-to-face meeting
Summary: discussed the project specification

Date: 01-11-2018
Supervision type: email
Summary: sent the draft specification and got the feedback

Date: 07-11-2018
Supervision type: face-to-face meeting
Summary: discussed the draft specification and follow-up arrangements

Date: 09-11-2018
Supervision type: face-to-face meeting
Summary: discussed the draft specification

Date: 11-11-2018
Supervision type: email
Summary: sent the updated specification

Date: 12-11-2018
Supervision type: email
Summary: some corrections were made by the supervisor

Date: 12-11-2018
Supervision type: email
Summary: sent the final version of specification

Date: 13-11-2018
Supervision type: email
Summary: got the approval of the specification

Date: 29-11-2018
Supervision type: face-to-face meeting
Summary: discussed the current progress according to the GANTT chart and the methods for data collection

Date: 30-11-2018
Supervision type: email

Summary: discussed the methods of data collection.

Date: 09-12-2018
Supervision type: email
Summary: sent a work plan and got the feedback

Date: 18-12-2018
Supervision type: face-to-face meeting
Summary: discussed the data collection progress and follow-up arrangements

Date: 03-01-2019
Supervision type: email
Summary: ask for suggestions about research methods

Date: 31-01-2019
Supervision type: face-to-face meeting in TUAT
Summary: set goals for the following 3 weeks

Date: 04-02-2019
Supervision type: face-to-face meeting in TUAT
Summary: progress reporting to TUAT supervisor

Date: 13-02-2019
Supervision type: face-to-face meeting in TUAT
Summary: progress reporting to TUAT supervisor

Data: 15-02-2019
Supervision type: face-to-face meeting in TUAT
Summary: discussion of current progress with assistant professor

Date: 18-02-2019
Supervision type: face-to-face meeting in TUAT
Summary: progress reporting to TUAT supervisor

Date: 23-02-2019
Supervision type: email
Summary: Requested draft of background chapter and draft mid-term reports were sent late to supervisor. Supervisor commented draft mid-term report.

Date: 26-02-2019
Supervision type: face-to-face meeting in TUAT
Summary: rehearsal of mid-term oral examination

Date: 11-03-2019
Supervision type: presentation in TUAT
Summary: introduced the project to lab members and received feedback

Date: 26-03-2019
Supervision type: face-to-face meeting in TUAT
Summary: explained the progress of the project, received feedback for finished work and suggestions for the next step

Date: 03-04-2019
Supervision type: face-to-face meeting in TUAT

Summary: progress check

Date: 11-04-2019
Supervision type: face-to-face meeting in TUAT
Summary: mock viva

Date: 17-04-2019
Supervision type: email
Summary: feedback of draft report

Date: 21-04-2019
Supervision type: email
Summary: comments on experiment setting and results

Date: 22-04-2019
Supervision type: face-to-face meeting in TUAT
Summary: discussed about the experiment setting

# Risk Assessment

| Description of Risk | Description of Impact | Likelihood rating | Impact rating | Preventative actions |
|---|---|---|---|---|
| Microsoft close its service | Facial expression analysis cannot work | 1 | 5 | Prepare a back-up version of system using other analysis tools |
| The program has crashed | Have to restart the system | 2 | 3 | Close other unrelated applications |
| The robot has broken | The HRI system cannot be applied properly | 3 | 3 | Take care when booting, shutting down or moving the robot |
| The computer and robot are not in the same network | The robot cannot follow the instructions sent from computer | 3 | 3 | Check the network connection before execution |
| Webcam has been broken | System cannot analyse facial expression in real-time | 2 | 5 | Check both the computer's webcam and external webcam |
| Network lost | Facial expression analysis cannot work | 3 | 5 | Check the connection before execution; prepare hotspot Wi-Fi with cell phone |

## Environmental Impact Assessment

As for the hardware, the project uses Pepper robot. It is a complicated device, which takes a large amount of resource in manufacture. However, it will be recycled and continually used in other projects. Also, to identify the state of students in real-time, the webcam has to be kept open. The robot, the webcam, and the computer all cause consumption of electric, which requires the burning of fuel.

For the software, the code is executed on the computer, which also consumes energy. However, since the project applies cloud computing service for facial expression recognition, the program can run without high-performance computing, so that a certain amount of energy has been saved.