

Union College

Union | Digital Works

Honors Theses

Student Work

6-2022

Impact of Movements on Facial Expression Recognition

Zhebin Yin

Union College - Schenectady, NY

Follow this and additional works at: <https://digitalworks.union.edu/theses>



Part of the [Artificial Intelligence and Robotics Commons](#)

Recommended Citation

Yin, Zhebin, "Impact of Movements on Facial Expression Recognition" (2022). *Honors Theses*. 2578.
<https://digitalworks.union.edu/theses/2578>

This Open Access is brought to you for free and open access by the Student Work at Union | Digital Works. It has been accepted for inclusion in Honors Theses by an authorized administrator of Union | Digital Works. For more information, please contact digitalworks@union.edu.

Impact of Movements on Facial Expression Recognition

By

Zhebin Yin

* * * * *

Submitted in partial fulfillment
of the requirements for
Honors in the Department of Computer Science

UNION COLLEGE

May, 2022

Abstract

YIN, ZHEBIN Impact of Movements on Facial Expression Recognition. Department of Computer Science, May, 2022.

ADVISOR: Webb, Nick

The ability to recognize human emotions can be a useful skill for robots. Emotion recognition can help robots understand our responses to robot movements and actions. Human emotions can be recognized through facial expressions. Facial Expression Recognition (FER) is a well-established research area, however, the majority of prior research is based on static datasets of images. With robots often the subject is moving, the robot is moving, or both. The purpose of this research is to determine the impact of movement on facial expression recognition. We apply a pre-existing model for FER, which performs around 70.86% on a given collection of images. We experiment with three different conditions: No motion by subject or robot, motion by one of the human or robot, and finally both human and robot in motion. We then measure the impact on FER accuracy introduced by these movements. This research relates to Computer Vision, Machine Learning, and Human-Robot Interaction.

Contents

1	Introduction	1
2	Background and Related Work	2
2.1	FER2013	4
2.2	MobileNetV2	4
3	Methodology	6
3.1	Experiment Preparation	6
3.1.1	Face detection	6
3.2	Movement Experiment	7
3.2.1	Static Motion	8
3.2.2	Motions of either the robot or subject	8
3.2.3	Motions of both the robot and the subject	11
3.3	Pipeline	12
4	Result	13
5	Future Work and Discussion	15

List of Figures

1	Social Robot - VALERIE at Union College	3
2	Social Robot - VALERIE at Union College	3
3	Static happy emotion of subject 1 through using the pipeline discussed in methodology. . . .	4
4	MobileNetV2 Model Architecture	5
5	MobileNetV2 Model Architecture with input and output.	5
6	Model accuracy with/without face detection in terms of the distance.	7
7	Output image after using the face detection in OpenCV	7
8	Cropped image by using face detection in OpenCV.	7
9	Experiment table, eight experiments in total. Five columns: experiment number, robot move- ment, human movement, device needed, and graph indication.	9
10	Experiment 0: static motion, indication image.	9
11	Experiment 1: Human moves parallel to the robot, indication image.	10
12	Experiment 2: Human moves toward the robot, indication image.	10
13	Experiment 3: the robot moves parallel to the human subject, indication image.	10
14	Experiment 4: the robot moves toward the human subject, indication image.	11
15	Experiment 5: the robot and human moves parallel to each other in the same direction	11
16	Experiment 6: the robot and human moves parallel to each other in the opposite direction . .	12
17	Experiment 7: the robot and human moves toward each other.	12
18	Pipeline of Experiment.	13
19	Target 1: Happy emotion with static motion.	13
20	Target 1: Happy emotion – Human moves parallel to the robot.	13
21	Target 3: Happy emotion with static motion.	14
22	Target 3: Happy emotion – Human moves parallel to the robot.	14
23	Extracted Images by face detection of subject 1 and subject 3 with happy facial expression. .	14
24	FER model accuracy, number of image extracted by face detection, and number of images predicted correctly for eight different motion experiment.	15
25	The P-value of each movement experiment through using the significant test.	16

List of Tables

1 Introduction

Humans are frequently exposed to interactions with robots more than ever under this quickly-developing technological era. In malls, there are robots that guide you to the store you are looking for. In hotels, delivery robots are replacing hotel staff to deliver food and convenience items. Even at home, robots such as sweeping robots are assisting with cleaning and housework. By collaborating with robots, the quality of our daily lives can be considerably improved. However, because of robots' limitations in catching human emotions and desires, the robot can present an improper response while interacting with humans, which may cause inefficient Human-Robot Interaction (HRI). In order to achieve the smooth HRI, the understanding towards human emotion is crucial for a robot to execute a proper reaction. For example, if humans express fear through their faces while collaborating with the robot, by recognizing the emotion of fear, the robot would immediately stop operating HRI task. *Our research question was: How would human and robot movements impact emotional recognition detection.*

The most apparent and direct way for humans to express their emotions without language is through their facial gestures and expression. Human face movements process superior expressive ability and generate "one of the most powerful, versatile and natural means of communicating motivational and affective state" [4]. The expressions of emotions have been studied for a long time. In the nineteenth century, Charles Darwin based the majority of his work on emotional expressions and experience in psychology [1]. We use facial expressions to provide the cues of emotions, intention, and mindset during social interaction. In the 1900s, according to the research of Ekman and Friesen, there were six basic emotions that facial expressions could be categorized into: anger, disgust, fear, happiness, surprise, and sadness [5]. Besides that, neutral expression is also an important facial expression that we should not neglect. It has fewer stimuli than the other six emotions, but still takes over a majority of one's daily time. According to the study in 2003, the "facial expression constitutes 55% of the effect of a communicated message" [10]. Hence, to make robots interact more properly with humans, to some extent, the robot should be able to recognize human emotions through the face of its users.

There are a large number of previous works based on the dataset collected under controlled conditions (e.g. clear images with static facial expressions). By using the extended Cohn-Kanade (CK+) database, recognizing facial expressions—seven universal emotions in total—through machine learning techniques achieved a high accuracy [8]. CK+ is the laboratory-controlled database that contains a sequence of images that illustrate a shift from a neutral expression to extensively exaggerated expressions [9]. Moreover, by using the dataset FER2013 collected by Google, the existing facial expression recognition architectures (CNN, VGG, VGGNet, etc) accomplished the highest accuracy of 76.82%. The database was less controlled

compared with the CK+, but also in an ideal environment where exaggerated facial emotions images are clear and with static human faces at the center.

However, a lot of factors in the less optimal environment should be considered if the robot is interacting humans in the real world applications. The less optimal conditions include: the distance between robots and people, movements of both robots and targets, covers on the human's faces like masks, lights in the environment, angle of the camera on the robot, and so on. According to the research in 2002, movements and physical interactions are the key ingredients to acquire smooth HRI in the human community Kanda.

Furthermore, in order to use the FER learning architecture in the mobile platform (i.e. the robots or mobile device), a relatively low complexity model is desired with appropriate accuracy. Prior work has found that using the MobileNetV2 model could greatly reduce the complexity of the parameters, 5 times smaller in size compared to the "winning entry to the FER2013 competition" [2]. Despite this, MobileNetV2 maintained a decent accuracy of 70.86%, which is only 6% lower than the highest accuracy with the collective training dataset created online by google (FER2013) using 28,709 training samples and 3,589 testing samples. In this research, we extended the result of the prior work on MobileNetV2 to examine the influence of movements, both human and robots', on the performance of this architecture. In particular, we evaluated the performance decrease caused by the different movements of robots/cameras and the target people under the natural environment.

Three sets of movement experiments were done in this study: no motion, some motion, and robot and human motion. The goal of the experiments was to measure the performance decrease caused by the movements of the robot and the subject. We used the Social Robot - VALERIE (Fig. 1) at Union College for this experiment. We used VALERIE as an example to see the influence of robot movement that might cause the performance to decrease. There are two cameras on the Social Robot, one at the bottom and one at the top. The camera on the bottom has a higher resolution but the position isn't well situated for FER. Note, the two preexisting cameras on VALERIE had poor resolution, making it hard for it to capture the images I desired. Instead, we used the webcam from Logitech of 1280*720 resolution to capture clearer images. We have the webcam sitting on the top of VALERIE to capture the faces of the targets (Fig. 2).

2 Background and Related Work

There were a lot of studies done on image recognition that is useful for the FER task. The occurrence of the ImageNet and the Deep Convolutional Neural Networks (CNN) [3] contributed a lot to the FER system. In particular, the ImageNet is a useful source for visual recognition applications such as object recognition, image classification, and object localization [7]. Neural Networks (NN) have "revolutionized many areas



Figure 1: Social Robot - VALERIE at Union College



Figure 2: Social Robot - VALERIE at Union College

of machine intelligence, enabling superhuman accuracy for challenging image recognition tasks” [6]. CNN systems could be adapted to a wide range of the environmental situation, “such as occlusion, head pose, and illumination variation ” [2]. We are going to use these neural architectures. However, they rely on large

quantities of data. The FER 2013 dataset is the most popular one for researchers to use.

2.1 FER2013

This is a set of data consisting of 28,709 training samples and 3,589 testing samples. Each sample is a 48*48 pixel grayscale image of human faces. These facial expressions in the dataset are also difficult for humans to recognize, the average accuracy for humans recognizing emotion correctly “is reported to be 65.5%” [2]. Figure 2 below is an example of the images in this dataset.



Figure 3: Static happy emotion of subject 1 through using the pipeline discussed in methodology.

2.2 MobileNetV2

The demand for high accuracy by using the Convolutional CNN requires high computational resources. The FER is a complex task that needs to be applied based on the extracted features; the coating interfaces were determined using the shortest path algorithm [3]. CNN is particularly good at image classification and recognition due to its high accuracy. Unfortunately, such high computational resources can hardly be applied and embedded to the majority of the mobile platforms. A new neural network architecture called the MobileNetV2 that was developed in 2018 could solve the cost problem greatly. This neural network architecture, MobileNetV2, is specifically designed for the “mobile and resources constrained environments,” which is far more efficient and has less cost than the CNN model that the majority of the research on Facial Recognition used [6]. By using MobileNetV2, we could decrease the number of operations and the required memory, but with the descent accuracy [6]. The following figure is the brief model illustrations of MobileNetV2.

As shown in Figure 5, we have 7 emotions in total. The architecture of MobileNetV2 takes an input size of $96 \times 96 \times 3$. In this model, the convolutional layers have 24 filters, 17 inverted Residual block layers (using the structure shown in Fig.1 repeatedly), and another full convolutional layer with 1280 filters. For the output network, the model will first use the pooling layer to produce 1280 parameters as vectors, and

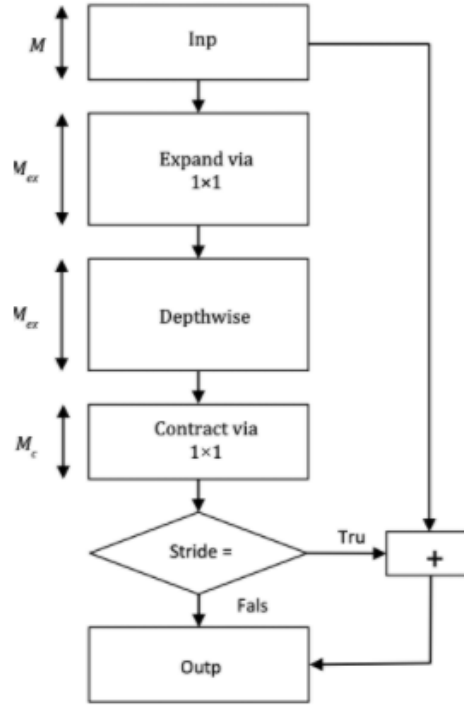


Figure 4: MobileNetV2 Model Architecture

then dense the layers to 7 neurons for each represented as an emotion. [2]

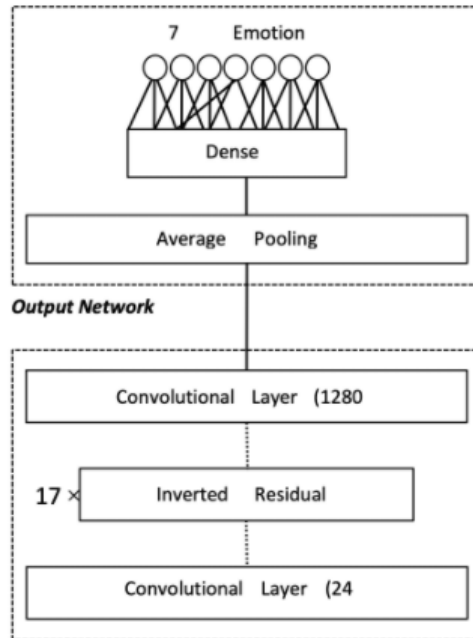


Figure 5: MobileNetV2 Model Architecture with input and output.

The MobileNetV2 architecture has been used to “feed into a fully connected output neural network of moderate size” [2]. They experimented with the performance of this architecture by using the FER2013

dataset with 35,887 faces via Google image search. By experimenting through using MobileNetV2 architecture, they found results where the accuracy was decreased by 0.3% compared with the accuracy of 71.16% achieved by the winning FER2013 entry. MobileNetV2 operates 2,340,423 parameters, which are 5 times smaller than the previous work with 12,000,000 parameters [2]. Furthermore, by using a 24 times less complex model, the best accuracy only dropped to 70.86%. This accuracy is still greater than the human baseline 65.5%, only decreased 5.36% than the best accuracy of the previous model [2].

However, this architecture the researchers had has only been used on the given dataset online with the collected image dataset. The accuracy of recognizing the emotions by using this model in the less optimal conditions remains unknown. Apart from that, the majority of the existing research on facial recognition either didn't apply it on the mobile platforms or mobile platforms that are with high-resolution cameras and advanced cores. There is little research that is applying facial recognition on the Robot, but more on devices like iPad and iPhone. In this experiment, we extended Professor Cotter's research by using this architecture in a real-time environment on the Social Robot. We used a camera with 1280*720 to capture the images in a real-time environment.

3 Methodology

3.1 Experiment Preparation

3.1.1 Face detection

OpenCV is a free video and image processing open-source library. It is used for image and video analysis, such as face detection, advanced robotic vision, photo editing, and so on. By using face detection in this library, we can easily detect whether the images have human faces or not. Face detection is using the Haar cascades, which is a "machine learning based approach where a cascade function is trained with a set of input data" [7]. In the OpenCV library, there are pre-trained classifiers for human face elements, like eyes, mouth, nose, amongst other features. There are a lot of other elements in the environment other than faces in images that could confuse the model architecture. Face detection in OpenCV, thus, is necessary to use in order to increase the performance of FER. Face detection can extract only the face caught in images so that the unrelated and confusing elements can be eliminated.

In preparation for our experiment, we examined the performance of the MobileNetV2 architecture with and without using the technique of face detection, considering different distances. In our experiment, the accuracy of using face detection before the architecture increased 31.5% overall with different distances (1, 2, and 3 meters) compared to the accuracy without using face detection. Thus, we had a precision of face

detection functions within 2 meters.

Distance	Model accuracy (without facial detection)	Model accuracy (with facial detection)
1m	19.25%	43%
2m	15%	58%
3m	15%	35%

Figure 6: Model accuracy with/without face detection in terms of the distance.

Below is the example of using face detection and an image of cropping the detected face.

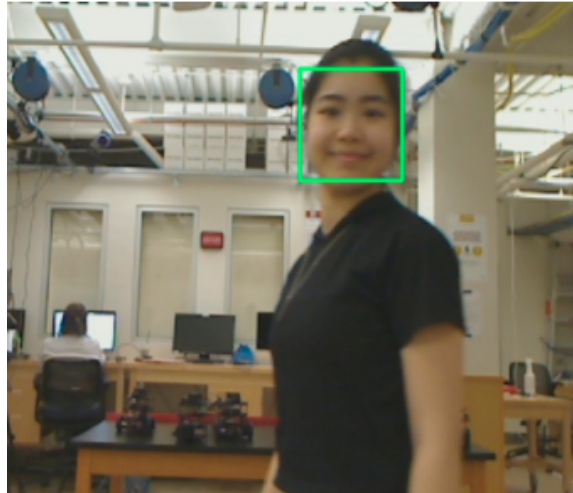


Figure 7: Output image after using the face detection in OpenCV .

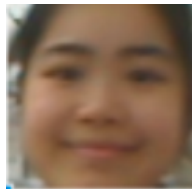


Figure 8: Cropped image by using face detection in OpenCV.

3.2 Movement Experiment

There were three sets of conditions: no motion of both robot and the subject (the controlled case), some motion of either robot or subject, and motion of both robot and subject at the same time. Among all the

different walking movements, walking parallel to and walking towards are the two most common movements to start and make the interaction between humans. Hence, walking parallel to and toward are two essential ways of movements that are necessary to examine. The human subject movement conditions are human moves parallel to and toward the robot. It is also necessary to test the movement of the robot since the vibration of robot movements may affect the image capture, thus weakening the performance of FER. We mimicked human subjects' movements as VALERIE's movements. The robot movement conditions are robot moves parallel to and toward the human subject. Furthermore, during HRI, humans and the robot are expected to move together. Thus, the movements for both the robot and the human subject are moving parallel to and toward each other at the same time.

We had three participants in total, and participants were those who understood this experiment and procedures in detail. The massive sample size was not required in this study since we didn't need to collect massive data to train the model, instead, we solely needed to observe what impact different movements would have on an individual, and across a small group.

Once the webcam was posed, the webcam will stay in its current position and angle. Subjects were asked to face the camera the entire time while the camera was recording. There was no rotational movement for the robot evolved in this experiment. Subjects were asked to walk slower than their normal speed (around 1.4 m/s). Figure 9 is the experiment table that showcases all of the controlled experiments. The experiment was done in the Crochet Lab at Union College. Note: all the datasets collected from our participants will remain private and protected.

3.2.1 Static Motion

Human and robot standstill at the set position without any movements. This is the controlled case where we compared cases with movement. The distance between robot and Subject is 1 meter. Figure 10 is the graph which indicates the experiment.

For each emotion, the subject needed to remain in the exaggerated emotion for 10 seconds. The webcam at the top of VALERIE would record for 10 seconds with frames per second (fps) of 30. Each subject needed to do seven different emotions.

3.2.2 Motions of either the robot or subject

Human moves while robot stays stationary

There were two parts of this experiment: human movement parallel to the robot and toward the robot. Due to the limited range of field, the webcam could capture, as well as the fps the webcam utilized to




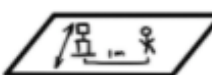
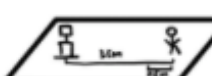
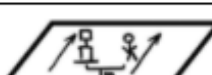
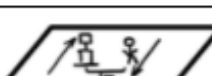
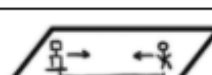
# Exp	Robot movement	Human movement	Device needed	Graph indication
0	No	No	A webcam will sit on the top of Social Robot	
1	No	Parallel	A webcam will sit on the top of Social Robot	
2	No	Toward	A webcam will sit on the top of Social Robot	
3	Parallel	No	A webcam will sit on the top of Social Robot	
4	Toward	No	A webcam will sit on the top of Social Robot	
5	Parallel ↑	Parallel ↑	A webcam will sit on the top of Social Robot	
6	Parallel ↑	Parallel ↓	A webcam will sit on the top of Social Robot	
7	Toward →	Toward ←	A webcam will sit on the top of Social Robot	

Figure 9: Experiment table, eight experiments in total. Five columns: experiment number, robot movement, human movement, device needed, and graph indication.



Figure 10: Experiment 0: static motion, indication image.

capture as many image messages as possible, subjects were required to walk slower than their normal speed (around 1.4 m/s). While the subject is moving, the subject will always look at the camera. The robot will stay in the same position all the time without any movement (i.e. rotations and moves), the same as the webcam sitting on the robot.

For the human subject moves parallel to the robot, the distance between human and robot was 1 meter. the subject moved from the right of the webcam view to the left and vice versa as one round. The subject did one round for each emotion, seven in total. We made sure that the subject moved out of the view each time from left to right and right to left. Subjects moved along with the walking trail that is displayed on

the floor with the blue tap, around 2.5 meters in total. Figure 11 below indicates the movements the subject requires to accomplish.



Figure 11: Experiment 1: Human moves parallel to the robot, indication image.

In the experiment of the human participant moves toward the robot, participants moved from the starting point to the endpoint marked on the ground by the experimenter. The starting point was 3.5 meters away from the robot and the endpoint was 0.5 meters away from the robot. Subjects moved along the walking trail with the required facial expression on their face and always looked at the camera. In total, the subject moved seven times with seven different facial expressions the experimenter asked. Figure 12 is the image which indicates the movement.

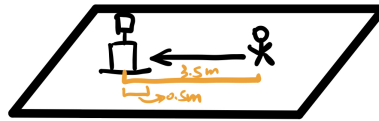


Figure 12: Experiment 2: Human moves toward the robot, indication image.

Robot moves while human stays stationary

There were two parts of this experiment: VALERIE moves parallel to the human subject and toward the human subject. The linear speed of VALERIE is 0.5 m/s. During the experiment, the human subject was required to always look at the webcam on the top of VALERIE.

For VALERIE moves parallel to the human subject case, the distance between VALERIE and the human subject was 1 meter. VALERIE moved from the subject's right to left and left to right with the camera toward the human at the fixed position on the top of VALERIE. The human subject faced and looked at the camera the entire time. VALERIE will move forward and backward with 2 meters each. The human subject was required to remain in the facial expression while VALERIE was moving and recording the subject's face. Figure 13 is the indicated image.



Figure 13: Experiment 3: the robot moves parallel to the human subject, indication image.

Another part of this experiment is the robot moves toward the human subject. VALERIE moved from 3.5 meters away from the participant until there was a 0.5-meter distance away from the participant. After finishing reaching the subject, the webcam stopped recording and VALERIE moved backward to the starting position. The participant stayed with the facial expression until VALERIE was 0.5 meters away from the subject. Figure 14 is the indicated image.

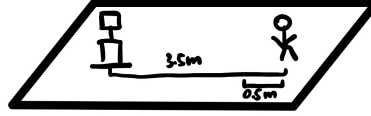


Figure 14: Experiment 4: the robot moves toward the human subject, indication image.

3.2.3 Motions of both the robot and the subject

Robot and the human subject moved together in this experiment. The participant was 1 meter away from the VALERIE. The linear speed of VALERIE was 0.5 m/s and the human subject was required to walk slower than their normal speed. There were three different cases: participant and VALERIE walk parallel to each other in the same direction, participant and VALERIE walk parallel to each other in a different direction, and participant and VALERIE walk toward each other.

The robot and human move parallel to each other in the same direction. The human subject was required to walk with VALERIE at the same time in the same direction. After the human subject made the required facial expression, the experimenter began recording and let VALERIE move forward. The participant looked at the camera the entire time and walked along with VALERIE at the same time until both of them reached the end position. The walking trail is 2 meters in total. After accomplishing the 2 meters walking, the experimenter then stopped the video recording and VALERIE will move backward to the starting point, waiting for the next round. Figure 15 is the indicated image.

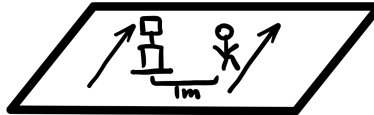


Figure 15: Experiment 5: the robot and human moves parallel to each other in the same direction

The robot and human move parallel to each other in the opposite direction. The participant and VALERIE started at a different position with a distance of 2 meters toward each other. The participant and VALERIE

then walked together at the same time. The participant and VALERIE walked past each other. After both of them reach the end position, the participant turned around and was ready with the facial expression while VALERIE stopped at the end position for 3 seconds to let the participant get ready. After 3 seconds, the participant and VALERIE will move toward each other parallelly at the same time again till both of them back to the starting position. Figure 16 is the indicated graph.

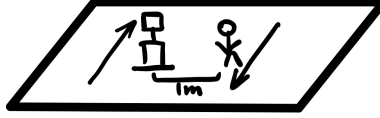


Figure 16: Experiment 6: the robot and human moves parallel to each other in the opposite direction

The robot and humans move towards each other. The participant and VALERIE started at a distance of 6 meters away from each other. The participant walked at the time the robot started to move forward. The participant stopped at the marked end position (2.5 meters away from the starting position) and the robot will stop after it moves 2.5 meters. That is, the participant and VALERIE ended until they were 1 meter away from each other. The participant was requested to stay in the facial expression while they are moving forward. Figure 17 is the indicated graph.

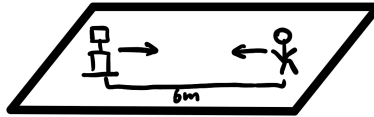


Figure 17: Experiment 7: the robot and human moves toward each other.

3.3 Pipeline

The webcam was sitting on VALERIE to record the video of each subject performing each emotion, controlled by ROS. After storing videos at the local computer disk, ROS was used to slice the video into image frames with 27 frames per second (fps). Then, by eliminating images without faces doing their assigned emotions, images were being uploaded to the Google Cloud. Face detection in OpenCV was used to extract faces from the image datasets. Then the extracted face images were fed into the weighted model which was previously trained. Through the trained model, the accuracy was being calculated and analyzed. Figure 18 is the graph of the full pipeline this study used. Figure 8 encapsulates the process of the experiment.

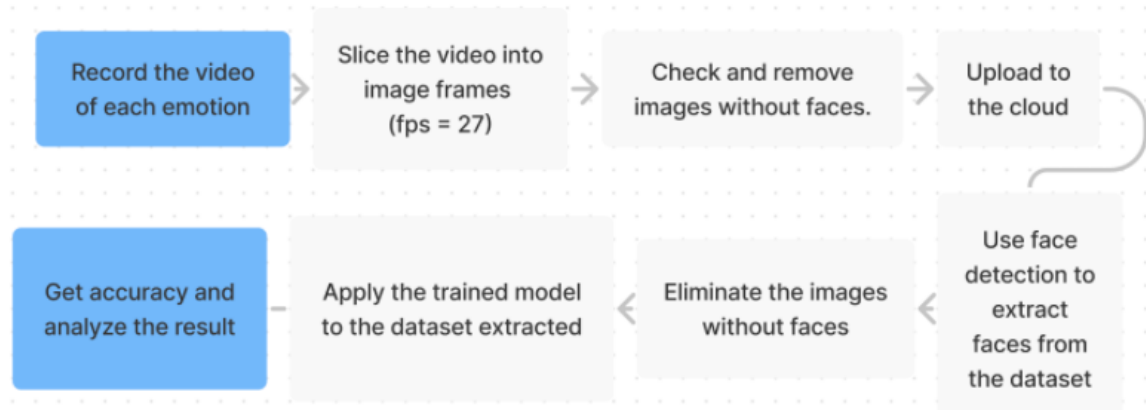


Figure 18: Pipeline of Experiment.

4 Result

Each of the participants was asked to pose exaggerated expressions of seven emotions: happy, sad, angry, disgust, fear, neutral, and surprise. The controlled factors were the number of emotions, the position of the webcam, the distance of the walking trail, and the speed of VALERIE. Figure 19 - Figure 22 are some examples of static motion and humans moving parallel to the robot by using the pipeline we discussed in the previous methodology section.

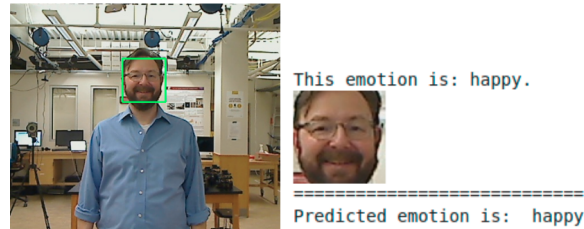


Figure 19: Target 1: Happy emotion with static motion.

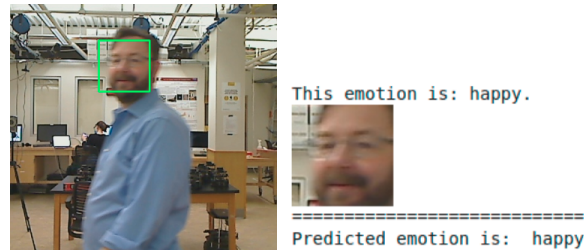


Figure 20: Target 1: Happy emotion – Human moves parallel to the robot.

Figure 23 is a set of images extracted from our subjects from eight experiment conditions. The images shown are all from the happy facial expression dataset collected. Although subjects were showing happy facial expressions, the images vary a lot depending on the frame sliced from the webcam captured. Some

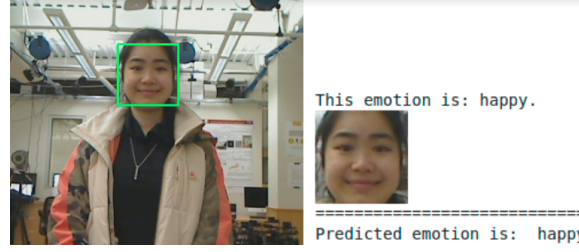


Figure 21: Target 3: Happy emotion with static motion.

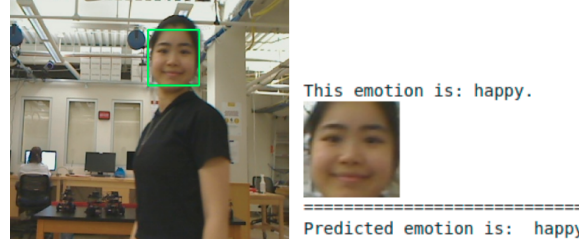


Figure 22: Target 3: Happy emotion – Human moves parallel to the robot.

extracted facial expression images are lighter than others. Some images are blurred because of the subject and robot's movements.



Figure 23: Extracted Images by face detection of subject 1 and subject 3 with happy facial expression.

Based on the dataset we collected from our three subjects, the table in Figure 24 below summarizes the overall accuracy of each experiment, images extracted by using face detection, and images predicted correctly by MobileNetV2 architecture. Overall, the accuracy of only human moves is better than the accuracy of only robot moves. The accuracy of human movements (HumanMotion1 and HumanMotion2) is 36.1% which is 6% higher than the accuracy of the robot movements (RobotMotion1 and RobotMotion2). Moreover, the parallel movement has less impact on the performance of FER than the forward movement with either the human subject or the robot moves in general. For human movements, the human subject's parallel movement accuracy is 20.6% higher than the forward movement. Similarly, for the robot movements, the robot's parallel movement accuracy is 13.1% higher than the robot's forward movement.

Other than that, for the condition of human and robot moving together, the situation in which the human subject and the robot move parallel in the opposite direction (HRMotion2) resulted in the lowest accuracy. The case where the robot and the human participant move parallel to each other in the same direction (HRMotion1) performed the best. In terms of the forward movement, the robot and the human subject moving forward to each other together contained the highest accuracy. Other than that, for the parallel movement, the robot stood still while the human participant was moving led to the highest accuracy.

Motions	FER_Accuracy	#ImageExtracted	#PredictedCorrectly
Static	46.60%	2868	1172
HumanMotion1	46.40%	2116	945
HumanMotion2	25.80%	1924	475
RobotMotion1	37.30%	4210	1948
RobotMotion2	24.20%	5785	1333
H&RMotion1	39.40%	3144	1276
H&RMotion2	28.80%	868	271
H&RMotion3	32.40%	2060	594

Figure 24: FER model accuracy, number of image extracted by face detection, and number of images predicted correctly for eight different motion experiment.

By doing the significant test, the figure 25 illustrates the p-value of 7 different movements. Compare the p-value of 0.05, the threshold, the p-values calculated by t-test are all greater than 0.05. Thus, based on the datasets we collected so far, there is no statistically significant difference between static and movement on the FER accuracy. That is, the movement of humans and the robot would not cause a statistically significant impact on the performance of FER.

However, based on the Fig. 24, the drop off between static and the movements of human and robot is severe, even if not significant. It is possible that with the increase number of participants, there will be a statistically significance. We concluded that there is a practical significant difference between the static and movements on the FER accuracy. That is, the movements of humans and robots would influence the emotional recognition detection.

5 Future Work and Discussion

Based on the experiment we made, we have the conclusion that movements will cause a severe drop off on the performance of FER. However, we didn't get a statistically significant difference due to possible factor of the small sample size. In the future, we will collect more data from larger sample size with the

Motions	p-value
Static	N/A
HumanMotion1	0.496
HumanMotion2	0.079
RobotMotion1	0.33
RobotMotion2	0.083
H&RMotion1	0.337
H&RMotion2	0.184
H&RMotion3	0.209

Figure 25: The P-value of each movement experiment through using the significant test.

pre-condition that all participants know nothing about the details of the experiments. We will also try to improve the model architecture to detect facial expressions more efficiently and accurately.

Furthermore, although we controlled a variety of conditions during the experiment, there were still a lot of factors that can't be controlled easily due to the limitations of the environment and the hardware. The uncontrolled factors include vibration of VALERIE during movements, variation of light from place to place in the environment, different status of the human subject while doing the facial expressions. During the experiment, we observed that VALERIE has the vibration when she is moving. She vibrates more when she is moving backward than moving forward. In the experiment, we involved the situation where VALERIE needed to move both forward and backward. The vibration of VALERIE would affect the quality of video recorded by the webcam which was sitting on the top of her, which may lead to the blur images that abate the performance of FER. Other than that, from the images extracted based on the face detection, we inspected that some frames are lighter than others. This may be caused by the difference of light in the environment and by the light adjustment ability of the webcam. It is possible that brighter images encompass higher contrast which may strengthen the performance of FER. Moreover, since the human subject is intelligent, alive, and easily distracted, the status of the subject and the elements around them while they were doing the facial expression may also have some effect either on their extent of doing different expressions or their ways of doing the same expressions.

We will examine the other different factors such as the lightness of the environment and vibrations of the robot. Other than that, it is also necessary to think about if the model could maintain a good accuracy in predicting less exaggerated expressions in a less constraint environment.

References

- [1] F. R. S. Etc Charles Darwin, M. A. *The Expression of the Emotions in Man and Animals*. London: 1872, Edinburgh review (1802) 137, no. 280 (1873)., September 1872.
- [2] Shane Cotter. Low complexity deep learning for mobile face expression recognition. Dec. 2019.
- [3] H. Gunes E. Sariyanidi and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37.
- [4] P. Ekman. *Emotion in the human face*. cambridge university press, cambridge.
- [5] P. Ekman and W. Friesen. The facial action coding system: A technique for the measurement of facial movement. 1978.
- [6] M. Sandler et al. Mobilenetv2: Inverted residuals and linear bottlenecks. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, page 258–64, 2003.
- [7] R.Socher L.-J.Li K.Li andL.FeiFei J.Deng, W.Dong. Face detection in 2 minutes using opencv python. *In Computer Vision and Pattern Recognition*, Medium, 2019.
- [8] et al. Khanzada, Amil. Facial expression recognition with deep learning. 2020.
- [9] S. Li and W. Deng. Deep facial expression recognition: A survey. 2018.
- [10] Philipp Michel and Rana El Kalioubi. "real time facial expression recognition in video using support vector machines.". *Proceedings of the 5th International Conference on Multimodal Interfaces*, page 4510–4520, 2018.