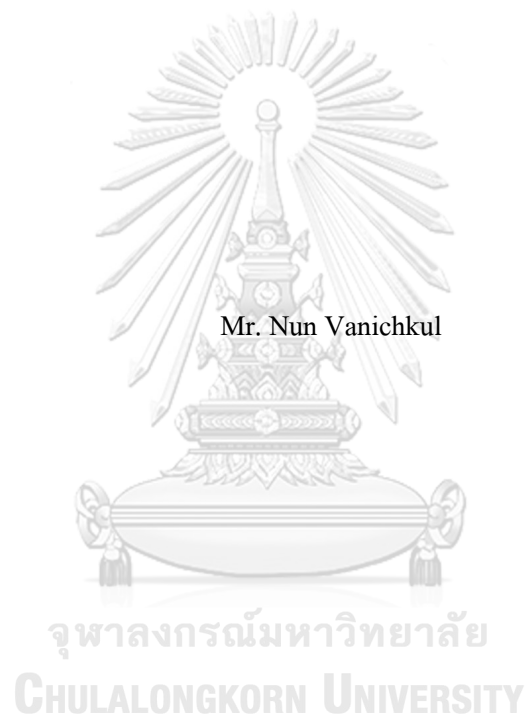


Confusion Detection from Facial Expression using Deep Neural Network



A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science

Department of Computer Engineering

FACULTY OF ENGINEERING

Chulalongkorn University

Academic Year 2020

Copyright of Chulalongkorn University

การตรวจจับความงุนงงจากการแสดงออกบนใบหน้าโดยใช้โครงข่ายประสาทเทียมเชิงลึก



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2563
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

นันทน์ วานิชกุล : การตรวจจับความงุนงงจากการแสดงออกบนใบหน้าโดยใช้
 โครงข่ายประสาทเทียมเชิงลึก. (Confusion Detection from Facial Expression using
 Deep Neural Network) อ.ที่ปรึกษาหลัก : รศ. ดร.ธนารัตน์ ชลิตาพงศ์

ความงุนงงเป็นอารมณ์ซึ่งถูกสังเกตได้บ่อยที่สุดในชีวิตประจำวัน และสามารถส่งผล
 อย่างมากต่อประสิทธิภาพและประสิทธิผลของการสื่อสาร โดยเฉพาะในการเรียนการสอน การ
 ตรวจจับความงุนงงจากผู้เรียนและแก้ไขได้อย่างทันเวลานั้นมีความสำคัญต่อความสำเร็จในการ
 สอนมาก งานวิจัยเกี่ยวกับการรับรู้จากการแสดงออกทางสีหน้าส่วนใหญ่เน้นไปที่การตรวจจับ
 เฉพาะหกอารมณ์พื้นฐานได้แก่ มีความสุข เศร้า โกรธ กลัว รังเกียจ ประหลาดใจ ถึงแม้เมื่อเร็วๆนี้
 โจทย์การตรวจจับความงุนงงจะได้รับความสนใจมากขึ้นจากนักวิจัยแล้วก็ตาม แต่การวิเคราะห์
 ทั้งข้อมูลเชิงพื้นที่และข้อมูลเชิงเวลาจากชุดข้อมูลที่มีปริมาณเพียงพอ นั้นยังคงขาดแคลนอยู่ ใน
 งานวิจัยนี้เรานำเสนอ โครงข่ายเชิงพื้นที่และเวลาสำหรับตรวจจับความงุนงงจากวิดีโอที่
 เรียนรู้จากชุดข้อมูล BAUM-1 ซึ่งเป็นชุดข้อมูลวิดีโอทัศนศาสตร์ขนาดใหญ่ที่สุดเท่าที่เราทราบว่ามี
 ระบุความงุนงง โดยโครงข่ายนั้นประกอบด้วย ResNet-18 Convolutional Neural Network
 (CNN) และ Long-Short Term Memory (LSTM) recurrent neural network (RNN) จากการนำ
 โครงข่ายประสาทเทียมเชิงลึกทั้งสองนี้มาเรียงต่อกัน ทำให้ได้ผลลัพธ์ที่แม่นยำถึง 73% บนชุด
 ข้อมูล BAUM-1s ซึ่งมากกว่าแบบจำลองสำหรับเปรียบเทียบซึ่งใช้โครงสร้าง LSTM ที่ 67% และ
 เราได้ทดสอบแบบจำลองที่นำเสนอกับชุดข้อมูลวิดีโอทัศนศาสตร์ความงุนงงที่รวบรวมจากการบันทึกภาพ
 ใบหน้าในระหว่างรับชมวิดีโอที่นำงุนงงของผู้เข้าร่วมการทดลองจำนวน 15 คนใน
 สภาพแวดล้อมที่ไม่มีการควบคุม โดยแบบจำลองสามารถทำนาย 1 ตัวอย่างซึ่งประกอบด้วย
 รูปภาพใบหน้าที่ต่อเนื่องกันจำนวน 30 รูปได้ภายในเวลา 0.04 วินาที และได้ความแม่นยำที่ 66%

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์
 ปีการศึกษา 2563

ลายมือชื่อนิติต
 ลายมือชื่อ อ.ที่ปรึกษาหลัก

6270138921 : MAJOR COMPUTER SCIENCE

KEYWORD: Facial Expression, Confusion Detection, Emotion, Deep Neural Network

Nun Vanichkul : Confusion Detection from Facial Expression using Deep Neural Network. Advisor: Assoc. Prof. THANARAT CHALIDABHONGSE, Ph.D.

Confusion is the most frequently observed emotion in daily life and can greatly affect the effectiveness and efficiency of communication. Detecting the confusion from learners and resolving timely is critical for achieving successful teaching in education. Most Facial Expression Recognition (FER) research works focus only on detecting six basic emotions: happiness, sadness, anger, fear, disgust, and surprise. Even though the confusion detection problem gains more attention from researchers recently, analysis of both spatial and temporal information with sufficient data is still short. In this study, we present a spatial-temporal network for confusion detection on video level which was trained on BAUM-1 database, as far as we know, this is the largest public video dataset which confusion is labeled. The model includes ResNet-18 Convolutional Neural Network (CNN), and Long-Short Term Memory (LSTM) recurrent neural network (RNN). By cascading these two deep learning structures, our method yields 73% accuracy which outperforms the baseline LSTM network that yields 67% on the same BAUM-1s dataset. We also test our proposed method with our confusion video dataset which was collected by recording 15 participants under uncontrolled environment. The model was able to predict 1 instance of 30 consecutive facial images within 0.04 seconds and got 66% of accuracy.

Field of Study: Computer Science

Student's Signature

Academic Year: 2020

Advisor's Signature

ACKNOWLEDGEMENTS

First, I would like to express my sincerest gratitude to my advisor Assoc. Prof. Dr. Thanarat Chalidabhongse and my technical consultant Dr. Thananop Kobchaisawat for their tireless efforts, support, and guidance throughout the entire course of my study.

I also would like to thank the member of my thesis committee, Asst. Prof. Dr. Peerapon Vateekul, Dr. Ekapol Chuangsuwanich, and Dr. Supakorn Siddhichai, for giving critical reviews of this work and for their advice on my proposal.

I would like to thank all of the participants who kindly taking their time to share information which crucial for this study, and also thank for my family and Suwara Boonpakorn for supporting and encouraging me from the very start of my master degree life.

Nun Vanichkul

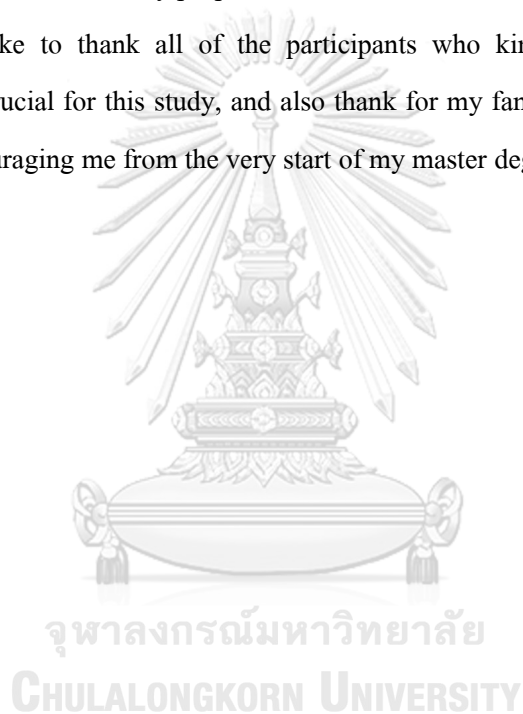


TABLE OF CONTENTS

	Page
ABSTRACT (THAI).....	iii
ABSTRACT (ENGLISH).....	iv
ACKNOWLEDGEMENTS.....	v
TABLE OF CONTENTS.....	vi
LIST OF TABLES.....	1
LIST OF FIGURES.....	2
CHAPTER 1 INTRODUCTION.....	4
1.1 Motivation and Problem Statement.....	4
1.2 Objective.....	5
1.3 Scope.....	5
1.4 Contributions.....	5
CHAPTER 2 RELATED THEORY AND LITERATURE REVIEW.....	7
2.1 Six basic emotions.....	7
2.2 Facial Detection & Facial Landmark.....	7
2.3 Facial Action Coding System (FACS).....	8
2.4 Psychological and Facial Expression of Confusion.....	9
2.5 Universality of confusion expression and recognition.....	11
2.6 Convolutional Neural Network (CNN).....	12
2.7 Residual Network.....	13
2.8 Long-Short Term Memory Network.....	13
2.9 Literature survey.....	15

CHAPTER 3 SPATIAL-TEMPORAL NETWORK FOR CONFUSION DETECTION	21
3.1 BAUM-1 database	21
3.1.1 Data acquisition	21
3.1.2 Annotation	21
3.1.3 Data preprocessing	22
3.2 Our CUPIC-Confusion videos dataset	24
3.2.1 Data acquisition	24
3.2.2 Annotation	26
3.3 Baseline LSTM confusion detection model	26
3.3.1 FAC extraction	27
3.3.2 Preprocessing	29
3.3.3 Training	29
3.4 Spatial-Temporal Network model	30
3.4.1 Preprocessing	30
3.4.2 Training Spatial-Temporal network and hyperparameters tuning	32
CHAPTER 4 TESTING PROCEDURE AND EXPERIMENTAL RESULT	38
4.1 Performance Measurements	38
4.2 Comparative evaluation using BAUM-1 database	40
4.3 Comparative evaluation using our CUPIC-Confusion video dataset	46
4.3.1 Testing procedure	46
4.3.2 Testing result	48
CHAPTER 5 CONCLUSION AND FUTURE WORKS	50
5.1 Conclusion	50
5.2 Recommendations for future works	51

REFERENCES53

VITA57



จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

LIST OF TABLES

Table 1 Single action units (AU) in the Facial Action Coding System [6].	9
Table 2 Basic Facial Movement Listed as Characteristic of Reports of Confusion-Puzzlement, Think-Concentrate, and Worry [16].	11
Table 3 Summary of BAUM-a after splitting the original video into frames.	23
Table 4 Summary of BAUM-1s after splitting the original video into frames.	24
Table 5 List of AUs in Openface 2.0 and Borges et al. [2].	28
Table 6 Testing results from the baseline and the proposed model on every measurement.	42
Table 7 Confusion Matrix of the baseline model on BAUM-1a dataset	43
Table 8 Confusion Matrix of the proposed model on BAUM-1a dataset.	44
Table 9 Performance of the proposed model on the confusion video dataset	48

LIST OF FIGURES

Figure 1 The 68 and 51 mark-ups for annotation [5].	8
Figure 2 Two-dimensional appraisal space for interest, confusion, and surprise [11].	10
Figure 3 Action units of expressing confusion and example facial image.	12
Figure 4 Residual learning: a building block [19].	13
Figure 5 Facial muscles [21].	16
Figure 6 Overall structure of Joint Fine-Tuning method [22].	17
Figure 7 Spatial-Temporal Network for facial expression recognition [23].	18
Figure 8 LSTM model proposed by Borges et al. [2]	20
Figure 9 Example of the frame that facial detector cannot detect the face because participant raises their hand to cover their face.	22
Figure 10 The samples of facial images from BAUM-1s labeled as Anger, Boredom, Bothered,	24
Figure 11 Sample from confusion video dataset, include a variety of lighting conditions, camera angle, sex, eyeglasses, and video quality.	26
Figure 12 Recreated confusion detection baseline model.	27
Figure 13 Loss and Accuracy of the training (orange) and validation (blue) dataset of training the baseline model.	30
Figure 14 8 image augmentation techniques compare to the original facial image.	31
Figure 15 Initial training process of the proposed spatial-temporal confusion detection network.	33
Figure 16 Loss and accuracy on training and validation datasets of ResNet-18 image classification model from the initial training process.	34
Figure 17 Loss and accuracy on training and validation dataset of spatial-temporal video classification model from the initial training process.	34

Figure 18 The final training process of the proposed spatial-temporal video classification network.....	35
Figure 19 Loss and accuracy of training ResNet-18 image classification model with the image-level dataset from merging the training and validation video-level datasets.	35
Figure 20 Structure of the proposed spatial-temporal network for confusion detection.	37
Figure 21 The loss and accuracy from training the final spatial-temporal confusion detection network.....	37
Figure 22 Confusion Matrix.....	38
Figure 23 ROC curve	39
Figure 24 Confusion matrix of the baseline and proposed model on their BAUM-1s testing dataset.....	41
Figure 25 ROC curve of the baseline model (left) and proposed model (right) on BAUM-1s testing dataset.....	41
Figure 26 Confusion matrix of the baseline and proposed model on the BAUM-1a dataset.	42
Figure 27 ROC curve of the baseline model (left) and the proposed model (right) on the BAUM-1a dataset.....	42
Figure 28 FPR of the baseline model on BAUM-1a testing dataset.....	45
Figure 29 FPR of the proposed model on BAUM-1a testing dataset.	45
Figure 30 Testing procedure on the confusion video dataset.	47
Figure 31 Confusion matrix of the testing result on the confusion video dataset.....	48
Figure 32 Pearson correlation coefficient similarity of each emotion/mental states with confusion, computed from average activation intensity of 17 AUs.....	51

CHAPTER 1 INTRODUCTION

1.1 Motivation and Problem Statement

One of the emotions that are mostly found in human communication is confusion. It can happen in many cases such as people communicating with foreigners who use different languages, and engineering students studying Einstein's special relativity theory in their freshmen year. When confusion occurs, it makes people face a hard time trying to understand and may have to invest a lot of time to comprehend unfamiliar information. In some cases, if people can not turn their confusion to be understanding fast enough, they have to take action or make decisions based on incomplete understanding which might be a misunderstanding or lack of detail that causes another time and effort to correct later.

When it is a face-to-face conversation, humans can naturally recognize confusion from each other's reactions such as facial expressions, body language, or questions from the listener. This is a good condition for mitigating confusion because it is instantly noticed and resolved. But this natural way of conversation has several limitations. First, speakers must be able to observe their listeners. Second, the capability of recognizing other's confusion is limited to a small number of people in the meantime. When the communication is on a large scale like public speaking, speakers might be able to recognize some audience confusion, but they cannot precisely know whether a whole audience can grasp what they are describing or not. And third, the accuracy of confusion recognition is depending on personal experience.

To overcome all the limitations, a system that can detect confusion automatically is necessary. Especially nowadays, people use various telecommunication technologies, such as video conferencing or massive open online courses, more than ever before. Confusion detection will increase the efficiency of communication by letting speakers know when their audience gets confused and be able to respond more suitably with the level of audience understanding which is essential for educational purposes.

Recently, several methods have been proposed to detect confusion automatically. Ranging from measuring electroencephalogram from the brain, electromyography from facial muscles, and Facial Expression Recognition (FER) from images. With the current technology, the first and second methods still need a specialist to set up and install sensors on the subject's body. While FER uses only a camera to record the subject's appearance. Although it seems more

convenient than others, this method still has many more aspects to study. Most research on this field focuses on classifying six basic emotions in which confusion is not included. Even some research begins to study this topic. But most of them do not take full advantage of deep learning techniques or using too limited amounts of data. These led to our motivation to find a method that can detect confusion from video on real-life implementation.

1.2 Objective

The objective of this study is to develop a method for automatic confusion detection from videos using a deep learning approach.

1.3 Scope

- Confusion in this study means the appearance of human facial expressions.
- Thinking state is considered as a part of confusion since we assume that people do not fully understand while they are still thinking.
- Confusion in this study only covers one that is occurred by comprehending new concepts or information with normal health conditions, not including confusion that is caused by injury, illness, drugs, or loss of situation awareness.
- The proposed method is trained and tested on videos from BAUM-1 Database which were recorded by [1].
- The performance of our proposed method is compared with the video-based LSTM network from [2] as a baseline of confusion detection from the video level.
- Since BAUM-1 Database was recorded under controlled environment, we collect our videos dataset by asking participants to record their faces while watching a confusing video in various places. As far as we know, it is the first dataset that confusion is labeled on videos under uncontrolled environment.

1.4 Contributions

1. The proposed spatial-temporal network that considers both spatial features of each frame and temporal relationship of them on video sequence which achieve comparable accuracy to the chosen recreated baseline on, as far as we know, the largest public video dataset that confusion is labeled. We also report the inference speed of the model which we consider to be important information for implementing the system in real-time.

2. Exploration from testing the proposed network with CUPIC-Confusion video dataset. Our dataset was collected under uncontrolled environment which is the most challenging and most realistic dataset that confusion detection has ever been studied before.



CHAPTER 2 RELATED THEORY AND LITERATURE REVIEW

2.1 Six basic emotions

Six basic emotions in [3] consist of Happiness, Sadness, Anger, Surprise, Disgust, and Fear. Facial expressions of these six emotions show evidence that they are universal. People who live in the preliterate culture and have minimal contact with foreigners can discriminate emotion from photographs of people from literate culture and vice versa. This supports the hypothesis that the association between facial muscular patterns and discrete emotion is universal.

2.2 Facial Detection & Facial Landmark

Before classifying human emotion from images, the crucial task that severely affects the accuracy of the whole process is facial detection, because emotion detection with deep learning needs a lot of training images. If the facial detection algorithm, which crops only the face region on the whole image, performs not well enough, it will limit the performance of the emotion detection algorithm by the quality of input image data. The classic and widely employed method is Haar feature-based cascade classifiers proposed in [4]. However, this task is still an active research topic, many challenges come when the images data are collected from an uncontrolled environment, for example, lighting condition, image quality, angle of camera and face, etc.

After the face region is detected, the important feature to be extracted is the position of each element on the face. Especially the region that is essential for emotional expressions such as eyes, eyebrows, and mouth. The comprehensive standard facial landmarks were proposed by Sagonas et al. [5]. As shown in Figure 1, the landmarks capture all parts of the eyebrow, eyes, nose, upper and lower lips which can be analyzed on each image as a static image-based emotion detection. Furthermore, tracking the position of the landmarks from consecutive frames gives us facial movement features that can be analyzed as sequence-based emotion detection which is more realistic data of expressing emotions than a still image.

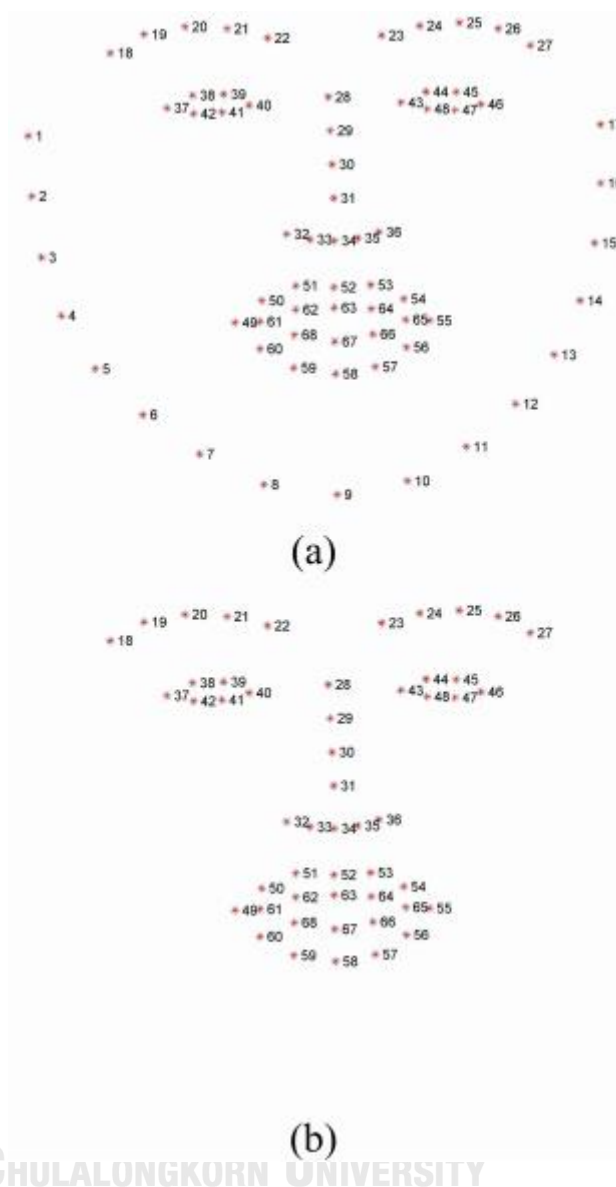


Figure 1 The 68 and 51 mark-ups for annotation [5].

2.3 Facial Action Coding System (FACS)

“The Facial Action Coding System is a comprehensive, anatomically based system for measuring all visually discernible facial movement” (Ekman & Rosenberg, 2005, p. 13 [6]). Facial activities including head pose, eye positions, and movements are labeled as unique numerical action units (AUs) that correspond to a single or a group of facial muscles on a five-point intensity scale. Many researchers apply these FACS as features for emotion recognition. The samples of AUs with corresponding facial muscles as shown in Table 1.

Table 1 Single action units (AU) in the Facial Action Coding System [6].

AU number	Descriptor	Muscular Basis
1.	Inner Brow Raiser	Frontalis, Pars Medialis
2.	Outer Brow Raiser	Frontalis, Pars Lateralis
4.	Brow Lowerer	Depressor Glabellae, Depressor Supercilli; Corrugator
5.	Upper Lid Raiser	Levator Palpebrae Superioris
6.	Cheek Raiser	Orbicularis Oculi, Pars Orbitalis
7.	Lid Tightener	Orbicularis Oculi, Pars Palebralis
9.	Nose Wrinkler	Levator Labii Superioris, Alaeque Nasi
10.	Upper Lip Raiser	Levator Labii Superioris, Caput Infraorbitalis
11.	Nasolabial Fold Deepener	Zygomatic Minor
12.	Lip Corner Puller	Zygomatic Major
13.	Cheek Puffer	Caninus
14.	Dimpler	Buccinator
15.	Lip Corner Depressor	Triangularis
16.	Lower Lip Depressor	Depressor Labii
17.	Chin Raiser	Mentalis
18.	Lip Puckerer	Incisivii Labii Superioris; Incisivii Labii Inferioris
20.	Lip Stretcher	Risorius
22.	Lip Funneler	Orbicularis Oris
23.	Lip Tightener	Orbicularis Oris
24.	Lip Pressor	Orbicularis Oris
25.	Lips Part	Depressor Labii, or Relaxation of Mentalis or Orbicularis Oris
26.	Jaw Drop	Massetter; Temporal and Internal Pterygoid Relaxed
27.	Mouth Stretch	Pterygoids; Digastric
28.	Lip Suck	Orbicularis Oris

2.4 Psychological and Facial Expression of Confusion

Confusion definition is defined by many psychological researchers, for example, confusion is a result from information that evokes more than one concept and thus creates cognitive conflict [7], “confusion is the feeling that the environment is giving insufficient or contradictory information” (Keltner and Shiota, 2003 [8]), and Ellsworth [9] speculated that confusion may stem from appraisals of uncertainty, and appraisal dimension in the Smith and Ellsworth [10]. Furthermore, Silvia [11] suggests that confusion and interest have different

positions in a two-dimension appraisal space: interesting things stem from appraisals of high novelty and high comprehensibility. Besides the definition, Confusion along with interest, surprise, and awe are classified into a family of knowledge emotions [12] [13] [14]. The knowledge emotions are caused by people's beliefs about their thoughts and knowledge, and these emotions stem from goals associated with learning.

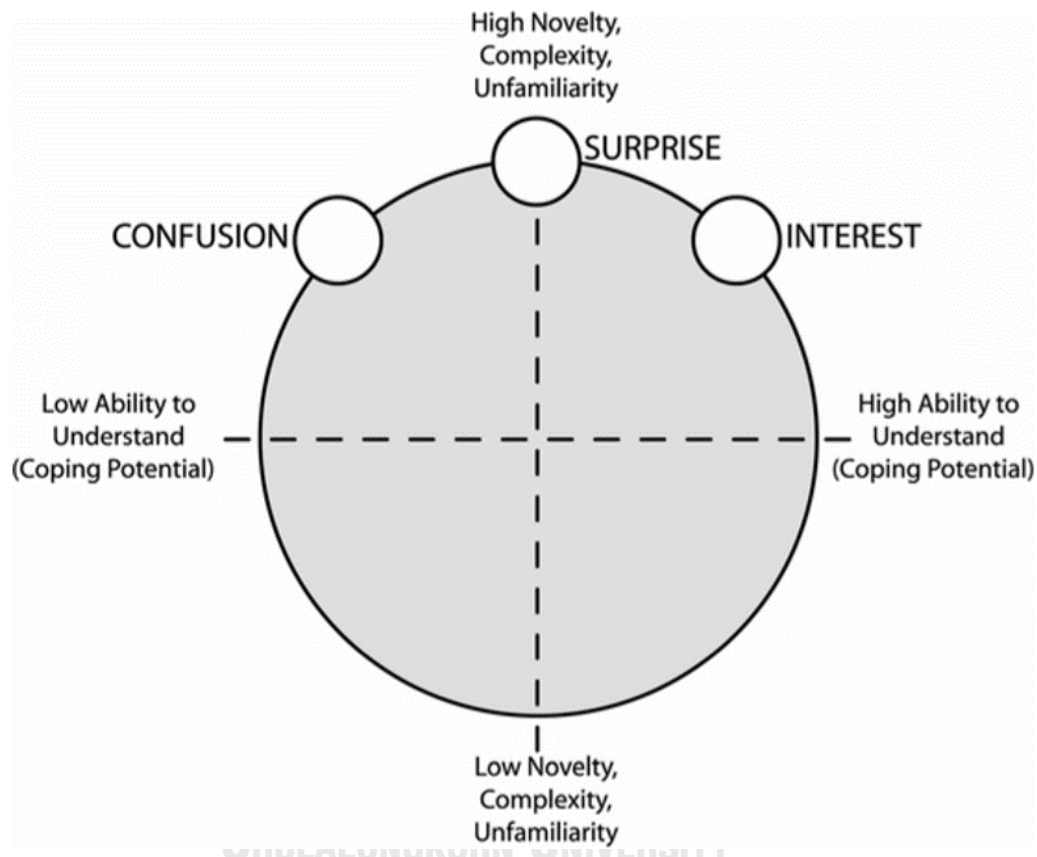


Figure 2 Two-dimensional appraisal space for interest, confusion, and surprise [11].

The first description of confusion's expression was discussed by Darwin [15] in the context of barriers to cognitive goals: "A man may be absorbed in the deepest thought, and his brow will remain smooth until he encounters some obstacle in his train of reasoning, or is interrupted by some disturbance, and then a frown passes like a shadow over his brow" (p. 223). After that, Rozin & Cohen [16] discovered from their study that confusion is the most frequent asymmetric emotion observed. Moreover, they also reported confusion's facial movements along with worry and concentration. From Table 2, most of the confusion's expression is from the eyes area which supports Durso et al. [17] experimental result from confusion detection with EMG.

Table 2 Basic Facial Movement Listed as Characteristic of Reports of Confusion-Puzzlement, Think-Concentrate, and Worry [16].

Facial action	Confusion		Worry		Concentrate	
	Symmet.	Asymm.	Symmet.	Asymm.	Symmet.	Asymm.
Eye area						
Eyebrow raise–arched	11	41		3	6	13
Eyebrow furrow–arched down–come together–knitted	10	8	6		11	
Eyebrow lower	2	6	2		3	4
Eye closed		1	1	1	1	7
Eye squint	7	13		2	6	9
Stare					6	
Eyebrows together–scrunch	1	2				
Frown	3	3	6	2	2	1
Eyes look up			1		3	5
Eyes wide open	3	1	5	2	1	
Mouth area						
Side mouth stretch						5
Lips–mouth raise	2	15		6		3
Lips purse–scrunch		5	3	1	5	3
Smile	5	9	3	2	4	6
Lip corner lower	1	3	1	4	2	
Lip press			4	1	3	1
Tongue protrude–move on teeth					1	5
Open	4		3	2	3	
Bite lip		2	4	16	5	6
Face scrunch		4	1			2

Note. Specific facial expressions are listed in this table only if reported at least eight times across the three nonstandard emotions or at least five times for any particular category of emotion and symmetry (Symmet.). Asymm. = asymmetry.

2.5 Universality of confusion expression and recognition

[18] performed an extensive study focused on exploring additional universal emotions more than the traditional six including happy, sad, anger, fear, disgust, and surprise. 5,500 video clips of expression were recorded from 120 participants from 5 cultures including China, India, Japan, Korea, and the USA. Each participant was told to express their emotion in response to 23 emotion stories read by a native of each culture. Facial expression video was chopped to contain only one expression per clip and coded with FACS to explore the pattern of emotional expression across cultures. From five cultures, the action units that were displayed at an above-chance rate across all cultures are called “international core sequence”. The international core sequence of confusion includes action units 4, 7, and 56 which represent Brown furrowed, eyelids narrowed, and head tilted respectively. Apart from the universality of emotion expression, the universality of emotion recognition across cultures is also studied by asking 453 participants, from China, Germany, India, Japan, Korea, New Zealand, Pakistan, Poland, Turkey, and America, to choose the expression that best fits the story from 5 choices including the correct facial image, 2 alternative facial images that were of the same valence with the correct choice, 1 most

physiologically similar well-studied emotion facial image, and “none of the above”. This study reveals that confusion expression from 5 cultures can be recognized and matched with correct emotion stories by participants from 10 cultures at an above-chance of accuracy. Therefore, this study concludes that confusion with weighted average facial expression recognition rate across ten cultures at 81% passes the 20% above-chance criteria for all cultures.


Emotion	Example photo	Action units	Physical description
Confusion		4+7+56	Brows furrowed, eyelids narrowed, head tilted

Figure 3 Action units of expressing confusion and example facial image.

2.6 Convolutional Neural Network (CNN)

One type of Artificial Neural Network (ANN) that is widely used in image classification is CNN which has a different structure from conventional ANN. The first layer after the input layer is the convolutional layer which each node is not connected to every node from the input image but only nodes or pixel values on a small area called receptive field. The weight of each node on a receptive field from the input image that is connected to each node on the convolutional layer can be represented as a small image with the same size as a receptive field called a filter. This filter performs a weighted sum operation with the area and slides to the next area repeatedly until complete with the whole area of the input image. A feature map is created as the outcome of a repeatedly weighted sum and is usually followed by a non-linear transformation function. Many different filters are applied to extract different feature maps from the input image. After that, the pooling layer extracts only strong features from the convolutional layer to decrease computational complexity. CNN generally stack several convolutional layers followed with pooling layers and convolutional layers again and so on before feeding the extracted features to fully connected layers and classifying image on the output layer. Compared to conventional ANN, CNN often reaches better accuracy with less computational complexity.

2.7 Residual Network

Residual network [19] or ResNet proposed to solve the degradation problem of conventional deep neural networks. The degradation problem occurs when the deep neural network has too many layers. Training accuracy of the model will be lower than the same structure network with fewer layers, whereas ResNet is not. This is because of a special structure of ResNet which is called Residual learning. Unlike conventional deep neural networks that every signal from one layer passes to the next layer, residual learning lets some signals skip some layers on shortcut connection, then both signals merge again on a deeper layer.

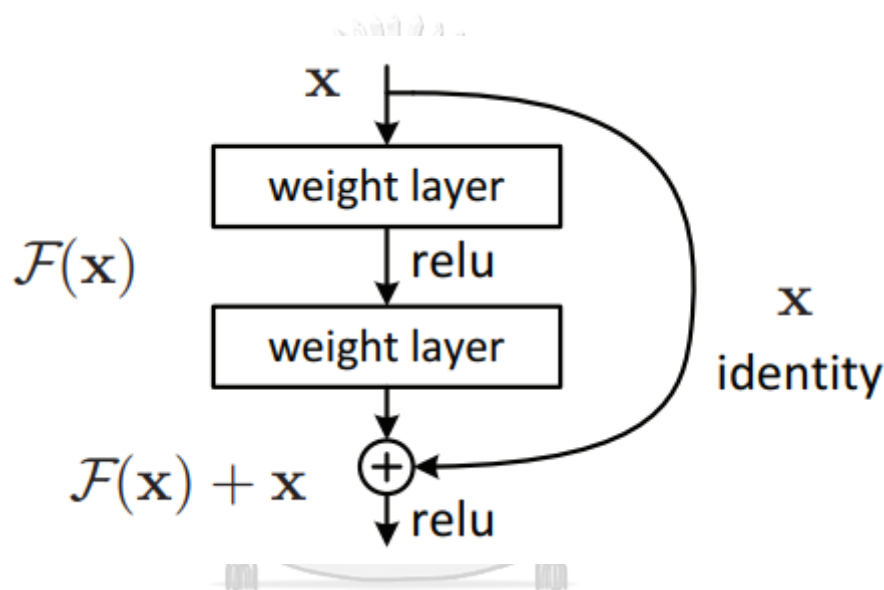


Figure 4 Residual learning: a building block [19].

Compared to the conventional deep neural network, the network with a shortcut connection takes advantage of a deeper layer of learning without losing training accuracy from the degradation problem.

2.8 Long-Short Term Memory Network

Long-Short Term Memory (LSTM) is one of the recurrent neural network variations. This deep learning structure can encode the temporal information from a sequence of input data by using the self-loop structure that can be represented by the LSTM cell. Each cell uses the cell state and hidden state of the previous cell and input data of the corresponding cell to predict the label of the next timestep. Three gate structures on each cell control the information that passes through the self-loop forward passing. Computation steps on each LSTM cell are described below.

Given that

W_{ij} : weight parameter which multiplies with matrix from i to calculate the output of j layer.

C_t : cell state matrix of timestep t

h_t : hidden state matrix of timestep t

f_t : the result of forgot gate of timestep t

i_t : the result of input gate of current cell t

g_t : candidate matrix of timestep t to update of the cell state C_t

O_t : output gate result of current cell t

b_i : bias term of the layer i

forgot gate uses a sigmoid function with h_{t-1} and x_t to calculate how much information of cell state from the previous cell C_{t-1} to forget.

$$f_t = \sigma(W_{hf}h_{t-1} + W_{if}x_t + b_f) \quad (1)$$

input gate with sigmoid layer decides how much information from candidate matrix g_t to be updated on cell state. The candidate matrix takes h_{t-1} and x_t as input of tanh function.

$$i_t = \sigma(W_{hi}h_{t-1} + W_{ii}x_t + b_i) \quad (2)$$

$$g_t = \tanh(W_{hg}h_{t-1} + W_{ig}x_t + b_g) \quad (3)$$

filter the information to forget by multiplying the cell state from the previous timestep with the result of forgot gate. Then add the result of the input gate that multiplied with the candidate matrix.

$$C_t = f_t * C_{t-1} + i_t * g_t \quad (4)$$

compute the output gate result O_t from the sigmoid function of the weighted sum of h_{t-1} and x_t then the hidden state can be calculated from the multiplication result of the O_t and the tanh function of cell state C_t .

$$O_t = \sigma(W_{ho}h_{t-1} + W_{io}x_t + b_o) \quad (5)$$

$$h_t = O_t * \tanh(C_t) \quad (6)$$

This model resolves the problem of recurrent neural networks that hardly learn the long-term dependency of input sequence data. Therefore, the model is suitable to apply with the problem that focuses on temporal information.

2.9 Literature survey

In this chapter, we review the literature relevant to our proposed study. We first take a brief overview of studies on various methods proposed for confusion detection. We then explore techniques on FER to classify six basic emotions. Finally, recent studies on confusion detection using the FER approach are described.

[17] proposed analyzing Electromyography (EMG) method. The assumption is that while the subject is expressing confusion, EMG which is an electrical signal from facial muscle will show patterns on its waveform that can be observed and used to differentiate confusion. From the experiments, they conclude that EMG is effective in detecting confusion and the corrugator supercilii (eyebrow) is the most diagnostic facial muscle while the zygomaticus major (cheek) is the least. Whereas training Gaussian Naïve Bayes classifier using Electroencephalography (EEG) data from Massive Open Online Course learner was proposed by Wang et al. (2013), which later, the classification method has been improved by Erwianda et al. [20] with XGBoost and Tree-Structured Parzen Estimator technique which reached accuracy at 87%.

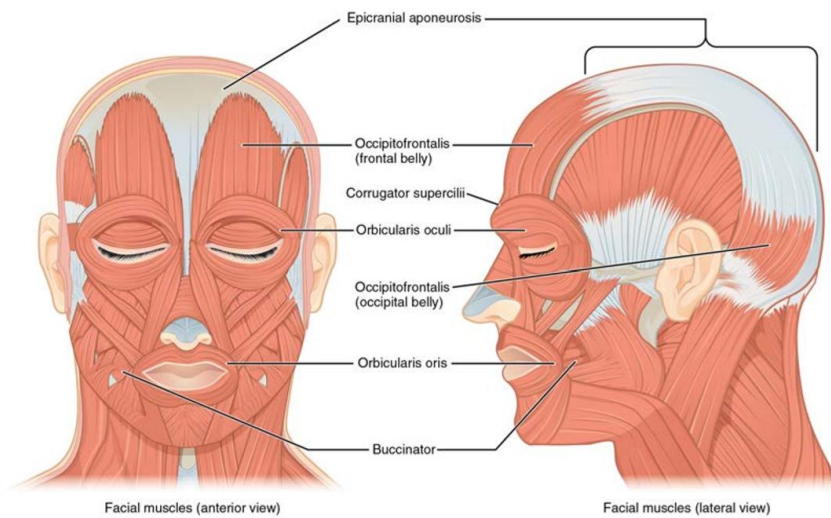


Figure 5 Facial muscles [21]¹.

As mentioned before, observing EMG and EEG requires attaching sensors to the subject's body. This attaching process needs a sensor itself and a specialist to set it up. On the other hand, cameras are a common tool that can be set up by anyone, and subjects do not have to confront an uncomfortable feeling from a sensor that is attached or installed on their body. Thus, on the aspect of setting up process complications and the subject's convenience, image processing seems more practical for large-scale implementation. Although most research on the Facial Expression Recognition (FER) field focuses on six basic emotions classification which consists of happiness, sadness, disgust, surprise, fear, and anger, the methodology and challenges are similar enough to mention.

Jung et al. [22] proposed to train two deep neural networks separately. The first network is the Deep Temporal Appearance Network (DTAN) which is used for capturing the temporal change of the input image sequence's appearance using 3D CNN structure. The second network is the Deep Temporal Geometry Network (DTGN) for capturing the temporal change of facial landmarks. Normalized 2D positions of each facial landmark are extracted as a 1D array from each image then sequences of arrays from consecutive frames are fed into a fully connected network with a softmax output layer. After separate training, both networks are combined by the joint fine-tuning method which freezes all hidden layers of both networks except the last one, then

¹ Download for free at <https://openstax.org/details/books/anatomy-and-physiology>

those two layers are additionally connected as input of a new fully connected layer with a softmax activation function. After that, the integrated network is trained by loss function calculated from both DTAN and DTGN cross-entropy loss functions. Only weights on the last hidden layer of each network and the new fully connected layer are tuned. Finally, the integrated model prediction came from the output of a new fully connected layer with a softmax activation function. This study got 97.25%, 81.46%, and 70.24% accuracy on CK+ (Lucey et al., 2010), Oulu-CASIA (Taini et al., 2008), and MMI (Valstar & Pantic, 2010), famous facial expression datasets, respectively.

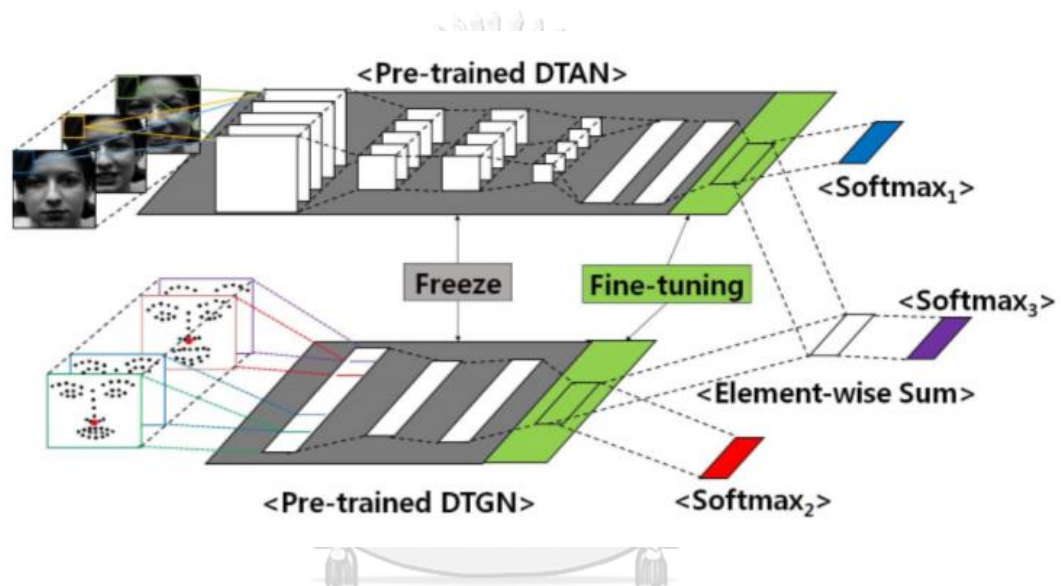


Figure 6 Overall structure of Joint Fine-Tuning method [22].

Zhang et al. [23] also proposed to train two deep neural networks separately and then combine them on the decision level too, but with a different methodology. The first network is a part-based hierarchical bidirectional recurrent neural network (PHRNN) that takes sequences of facial landmarks as input. Each facial landmark image is divided into four parts i.e. eyebrow, eyes, nose, and mouth. Then, Each Bidirectional Recurrent Neural Network (BRNN) subnet takes each part as input and gradually merges with each other until the upper layer receives all extracted signals from every part. After that, the signal goes through the LSTM-BRNN layer, fully connected layers, and softmax layer, respectively. Besides, the second network is Multi-Signal Convolutional Neural Network (MSCNN). The network takes grayscale static images as input and feeds them through four convolutional layers, one fully connected layer and a softmax layer, respectively. Two loss function is employed to train MSCNN, cross-entropy loss to learn an

expression recognition signal and loss function based on L2 norm to learn expression verification signal. The predicted probabilities of both networks are combined by a fusion function to provide final predicted probabilities.

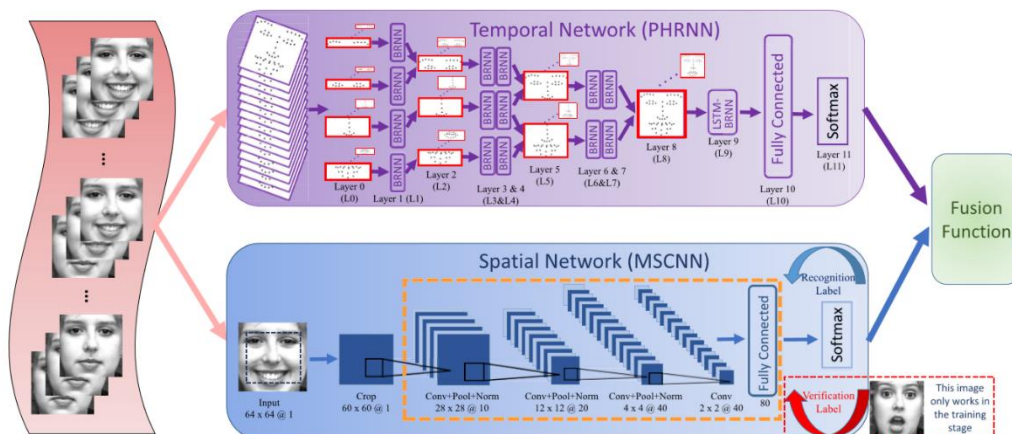


Figure 7 Spatial-Temporal Network for facial expression recognition [23].

Compared with six basic emotions, there are fewer datasets and algorithms that focus on detecting confusion [24], but some research begins to study this topic such as Zhalehpour et al. [1] Their study recorded both videos and audios of participants while they were watching stimuli images and videos. They performed a comprehensive experiment to create classifiers from visual features, audio features, and both. By using only visual features, their pipeline starts with cropping only part of the detected face region that is relevant to expression then extract Local Phase Quantization and Patterns of Oriented Edge Magnitudes to compute and select six peak frames from each video clip and assume that the peak frames contain the high intensity of expression. The images are then classified by Support Vector Machine (SVM) with the linear kernel. However, the accuracy of confusion detection was not explicitly mentioned but reported as average accuracy of classifying thirteen emotions and mental states, which are Anger, Disgust, Fear, Happiness, Sadness, Surprise, Boredom, Contempt, Unsure (include confused and undecided), Neutral, Thinking, Concentrating and Bothered, at 25.17%. This study performs well on data collection with decent quality and volume of data. 1,457 acted and spontaneous videos from 31 subjects were recorded and annotated with thirteen emotions/mental states by five annotators. This dataset greatly encourages other researchers to study FER in addition to six basic emotions.

Shi et al. [25] proposed the method to detect confusion specifically. Video data were recorded from 82 students while they were studying an online course and they also labeled when they were confused on their own clips. Then, the videos were separated into still images and processed with face detection, size normalization, rotated to be perpendicular, and cropped again to obtain the facial expression area. The pre-trained CNN VGG16 [26] was applied as a feature extractor and classification was done by the SVM classifier. Although this research specifically focuses on confusion detection, it still lags some aspects for implementing this method in real-world use. First, training and testing are performed with a dataset from the same source and collecting method. Therefore, there might be a lag of generalization which could be mitigated with data augmentation techniques. Second, the model is based on classifying each frame from video, while confusion is expressed as a combination of facial movements over a period of time. The feature from the still image might not as realistically as the video or sequence of consecutive frame features. Lastly, experimental results do not report inference speed which is essential for real-time implementation.

Borges et al. [2] proposed an image sequence-based classification method. After recorded videos of instructor-follower dyads engaging in a map directional task, the videos were analyzed at 15 frames per second with 2 seconds duration. The intensity level of activation from 20 action units of FACS was extracted from each frame. Thus, each instance representation is 20 action unit's activation intensity from 30 consecutive frames. Then, a neural network with 32 cells LSTM-layer, fully connected layer with 64 nodes, and softmax layer with 4 classes consisting of Positive, Negative, Neutral, and Confusion as demonstrated in Figure 2-4 was trained. The accuracy, recall, precision, and F1-score of the network are 87.07%, 84%, 78%, and 81% respectively. This research starts to train the model that detects confusion from features of image sequences which is more realistic for spontaneous confusion expression than still image classification, however, the intensity of 20 activation units might not conclude all features that are significant for confusion detection. Moreover, the fact that this study uses limited amounts of data (12 videos) raises the question of the generalization of this method.

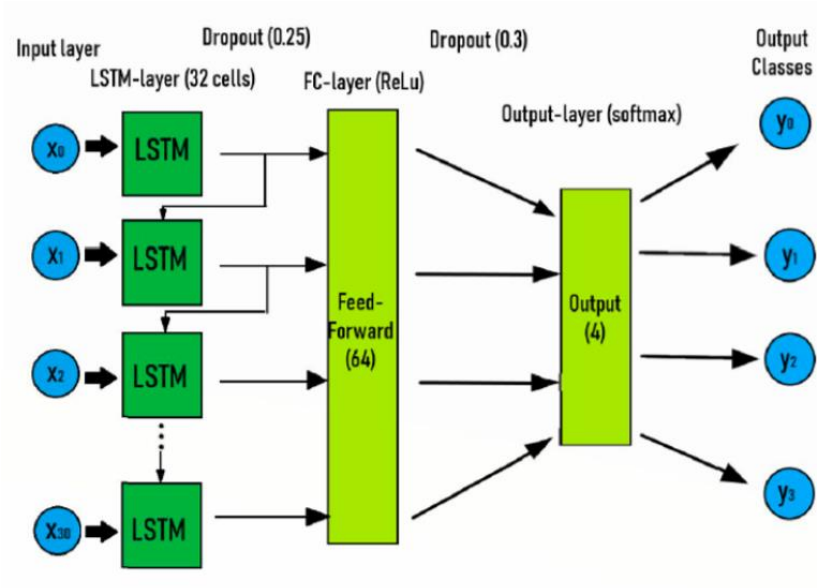


Figure 8 LSTM model proposed by Borges et al. [2]



CHAPTER 3 SPATIAL-TEMPORAL NETWORK FOR CONFUSION

DETECTION

In this chapter, we describe the details of structures and training procedures of our proposed method and baseline LSTM model. We also explain the details of datasets that were used in this study comprised of the BAUM-1 database, and our CUPIC-Confusion video dataset.

3.1 BAUM-1 database

3.1.1 Data acquisition

The BAUM-1 Database has 2 different parts of videos recorded from the same controlled studio; lighting condition, camera angle, video quality, and background are the same. In the first part, 31 participants were asked to utter several sentences with target emotions and imagining specific scenarios. The 8 target emotions of this part including happiness, sadness, fear, anger, disgust, unsure, boredom, and interest. This acted part which has 273 videos is called BAUM-1a. The second part is 1,134 videos of the same 31 participant's spontaneous expression from watching stimuli videos and images which are called BAUM-1s. the procedure of recording the BAUM-1s is that participants sit in front of a camera and a monitor. 29 stimuli images and videos are carefully chosen to elicit the desire emotion from the participants were shown. Target emotions and mental states of this part including unsure (confusion, undecided), concentrating, thinking, happiness, neutral, contempt, surprise, boredom, anger, sadness, disgust, fear, and bothered. There is 30 seconds gap between each image or video for participants to express their feeling in their own words. The total length of the record session is 50 minutes per person.

3.1.2 Annotation

The BAUM-1a videos are labeled with the target emotions. The raw 50 minutes video of each participant from BAUM-1s was chopped to be multiple short clips that contain only one expression per clip. Each clip was annotated by five annotators. For each clip, the annotators choose only the best fit from 13 emotions and mental states and rate the intensity of expression on the clip. The final label of each clip is decided by the majority voting over the five annotators. In our study, we focus only on the emotion and mental states that we group "Thinking" and "Unsure" as confusion and others are non-confusion.

3.1.3 Data preprocessing

BAUM-1 videos are separated into frames. Every video is recorded on the same 30 frames per second with 854 x 480 resolution, result in 28,599 frames from the BAUM-1a and 167,854 frames from the BAUM-1s. Then, we perform facial detection to crop only the facial region from every frame by using the facial detection algorithm adapted from CenterNet [27] which originally developed as an object detection algorithm, but we train the CenterNet on the WIDER FACE dataset in order to use it as a facial detector. From the facial detection step, we lost 8 frames on BAUM-1a and 64 frames on BAUM-1s because the facial detector can not detect the participant's face on that frame which cause by the expression of the participant occludes their face as the example in Figure 9.

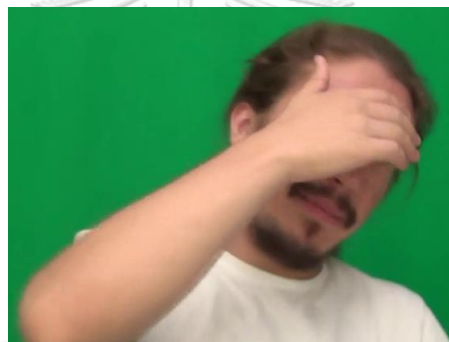


Figure 9 Example of the frame that facial detector cannot detect the face because participant raises their hand to cover their face.

After that, the file paths of facial images are mapped with their original video's label. 10 videos on BAUM-1a and 64 videos on BAUM-1s are unlabeled on the original labeling file cause to loss of every frame from those videos. Thus, a dataset of facial images with labeled emotion/mental states is created. Then, we create our target features "confusion" by grouping "Thinking" and "Unsure" from the original label as confusion and others as non-confusion. The number of subjects, video clips, facial images, and average clip duration (calculated from the number of facial images divide by 30 frames per second) of each emotion/mental states that we extracted from the original BAUM-1a and BAUM-1s videos is shown in Table 3 and Table 4 respectively. This frames level dataset is used to create a video level dataset which will be explained in the following sections.

Table 3 Summary of BAUM-1a after splitting the original video into frames.

Class	Emotion	#Subject	#Clip	#Image	Avg.Duration(s)
Non-Confusion		30	227	23,650	3.47
	Anger	28	42	3,848	3.05
	Boredom	23	26	2,630	3.37
	Disgust	29	34	2,591	2.54
	Fear	26	35	4,037	3.84
	Happiness	25	26	2,541	3.26
	Interest	27	28	2,750	3.27
	Sadness	24	36	5,253	4.86
Confusion		26	37	4,087	3.68
	Unsure	26	37	4,087	3.68
Grand Total		30	264	27,737	3.50

Table 4 Summary of BAUM-1s after splitting the original videos into frames.

Class	Emotion	#Subject	#Clip	#Image	Avg.Duration(s)
Non-Confusion		30	895	114,935	4.28
	Anger	21	56	6,088	3.62
	Boredom	11	22	1,961	2.97
	Bothered	23	91	12,249	4.49
	Concentrating	24	62	10,776	5.79
	Contempt	6	14	1,985	4.73
	Disgust	25	80	9,249	3.85
	Fear	16	37	3,273	2.95
	Happiness	30	173	19,852	3.83
	Neutral	29	185	26,740	4.82
	Sadness	25	134	19,242	4.79
	Surprise	19	41	3,520	2.86
Confusion		30	239	43,526	6.07
	Thinking	28	107	22,118	6.89
	Unsure	30	132	21,408	5.41
Grand Total		30	1,134	158,461	4.66



Figure 10 The samples of facial images from BAUM-1s labeled as Anger, Boredom, Bothered, Concentrating, Neutral, and Unsure, respectively.

3.2 Our CUPIC-Confusion videos dataset

3.2.1 Data acquisition

We want to test our proposed model with the most realistic data possible. Unfortunately, we can not find any public facial expression videos dataset that labels the confusion except the

BAUM-1 database that we use as the main dataset for this study. Therefore, we decided to collect our confusion video dataset to fulfill this need. We ask participants to send us a video of their faces while they are watching a confusing video. Participants are allowed to watch any video they want or watch the video about the hard logic puzzle “three gods riddle” that we prepare for them (access via <https://www.youtube.com/watch?v=LKvjIsyYng8>). Then, report whether they are confused from watching the video or not. Participants are also allowed to record the video from any environment with any device. Therefore, the lighting conditions, background, camera angle, and video quality are different for each video. Google form that we shared with the participants for data collection has mandatory fields including name, link to the google drive that the video is uploaded, self-report label (confusion or not), and the checkbox for information-sharing agreement. The optional fields on the google form are gender and birth date. Since recording video must be done on a device that can play video and record video simultaneously or play the video on one device and record on another device, and uploading video need a lot of internet usage, only 50 videos from 15 participants were acquired on a volunteer basis. Most of the participants chose to watch our confusing video which makes their video length around 5 minutes, but for others who watch other videos, their video length varies from shortest at 7 seconds to longest at 20 minutes. Among 50 videos we received, 48 of them were recorded on 30 frames per second (fps) frame rate, one on 28 fps, and the others on 24 fps. These frame rates, which are equivalent with the BAUM-1’s frame rate, make us confident to test the proposed method on the video that has the same frame rate as the training data.



Figure 11 Sample images from our dataset which includes a variety of lighting conditions, camera angles, genders, eyeglasses wearing, and video qualities.

3.2.2 Annotation

After collecting the videos from the participants with their self-reported labels, we annotate the time that confusion is recognized on every video as a pair of start and stop times. In some videos, the confusion is recognized more than one time, which in some videos, we can not recognize confusion even it is reported as confusion. We decided to annotate the part of the video as confusion only when the video has been reported as confusion and we also recognize it. After the videos are annotated with confusion time, we converted the confusion time to start and stop-frame by multiply the time with the video frame rate. This confusion interval dataset of each video will be used to test the proposed model as the most realistic testing dataset which will be described in Chapter 4.

3.3 Baseline LSTM confusion detection model

This baseline LSTM confusion detection model is recreated to align with the method proposed in [2]. We train this baseline model and our spatial-temporal network on the BAUM-1s

dataset to compare their result on detecting the confusion from the same controlled environment. For the input layer, we used one directional LSTM layer with 30 LSTM cells to take the action unit's activation intensities from the input sequence. Then, the hidden layers consist of the dropout with 25% probability, 64 nodes fully connected layer with Rectified Linear Unit function (ReLU) activation function, and the second dropout with 30% probability respectively. For the output layer, we have to adjust from 4 classes softmax layer from the original paper to be 1 node fully connected layer with sigmoid activation function to fit with the BAUM-1 dataset that we focus only on detecting confusion. The baseline model structure with x_i as activation intensity of i AU is demonstrated in Figure 12.

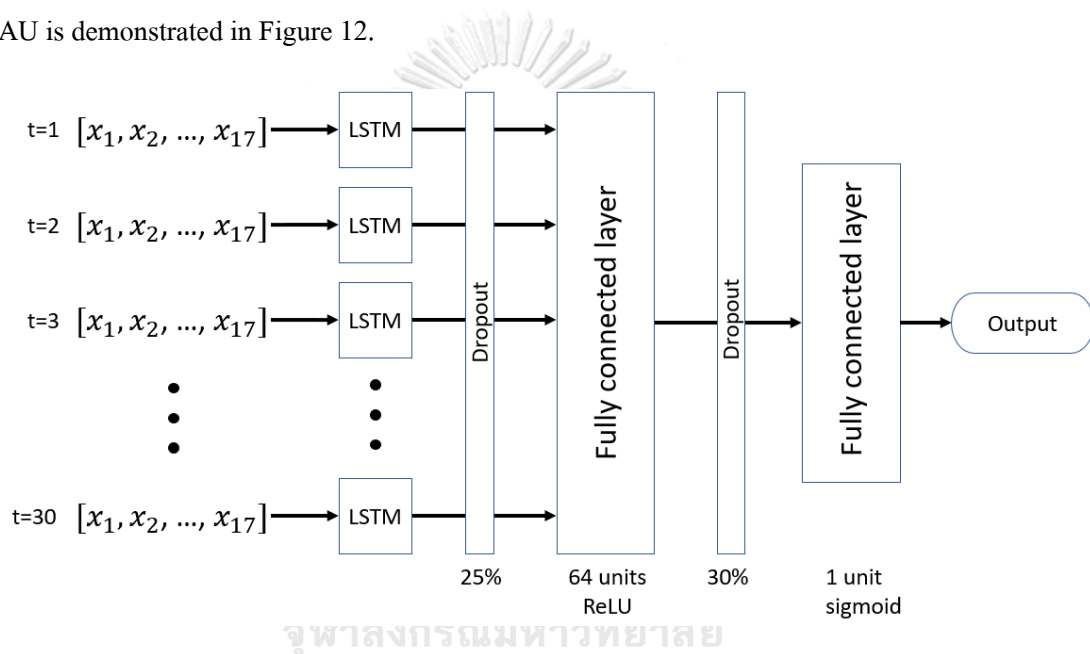


Figure 12 Recreated confusion detection baseline model.

3.3.1 FAC extraction

We use OpenFace 2.0 [28], facial behavior analyzing software, to extract action unit activation intensities from each frame of the BAUM-1 video. The OpenFace 2.0 uses the concatenation of dimensionality reduced Histogram Of Gradient orientations (HOGs) as a feature and used linear kernel SVM to detect AUs. The algorithm is trained on DISFA, SEMAINE, BP4D, UNBC-McMaster, Bosphorus, and FERA 2011 datasets. The presence and activation intensity of AUs are trained separately with a different number of AUs, the details compare with Noldus FaceReader that is used on the original baseline LSTM model as shown in Table 5. We wanted to extract both presence and activation intensity from every frame of BAUM-1 using OpenFace 2.0. Unfortunately, it was not easy to extract the features from every frame. For some

frames of BAUM-1s, we had to re-extract up to 5 times before the features were successfully extracted. Even after 5 times, we still failed to extract the features from 21,807 frames and had to leave them as unextractable frames. We faced the same problem on BAUM-1a which resulted in 5,350 unextractable frames. This problem was concerned us with the fairness of comparison between the baseline model and our proposed model which trained with all of the facial images.

Table 5 List of AUs in Openface 2.0 and Borges et al. [2].

AU No.	Description	Openface 2.0	Openface 2.0	Borges et al
		Intensity	Presence	
1	Inner Brow Raiser	✓	✓	✓
2	Outer Brow Raiser	✓	✓	✓
4	Brow Lowerer	✓	✓	✓
5	Upper Lid Raiser	✓	✓	✓
6	Cheek Raiser	✓	✓	✓
7	Lid Tightener	✓	✓	✓
9	Nose Wrinkler	✓	✓	✓
10	Upper Lip Raiser	✓	✓	✓
12	Lip Corner Pullers	✓	✓	✓
14	Dimpler	✓	✓	✓
15	Lip Corner Depressor	✓	✓	✓
17	Chin Raiser	✓	✓	✓
18	Lip Puckerer	✗	✗	✓
20	Lip Stretcher	✓	✓	✓
23	Lip Tightener	✓	✓	✓
24	Lip Pressor	✗	✗	✓
25	Lips Part	✓	✓	✓
26	Jaw Drop	✓	✓	✓
27	Mouth Stretch	✗	✗	✗
28	Lip suck	✗	✓	✗
43	Eyes Close	✗	✗	✓
45	Blink	✓	✓	✗

3.3.2 Preprocessing

After extract AUs intensity, we selected every other frame from the first two seconds of each 30 fps video. The activation intensities of each selected frame are stacked as an array of size 17×30 which is used as video representation on video level dataset. This resulted in each instance of our video dataset being represented by a sequence of 30 frames of 15 fps image sequence from the first 2-second duration of each video. Due to the facial detection and AUs extraction difficulty, some videos missed the activation intensity of some frames on the first 2 seconds duration. Therefore, we selected only the videos with at least 2 seconds duration and have no missing features during the interval to prevent some video instances that differ from other instances. Moreover, due to the imbalance between the numbers of confusion and non-confusion video instances on the dataset, we downsampled the non-confusion instances to have the same number as the confusion instances. Then, we split the video dataset into training, testing, and validation datasets with stratify method that preserves the ratio of confusion and non-confusion on the total dataset on each part. The splitting ratio is 80:10:10 which aligns with the original baseline paper. The number of video instances on the training, testing, and validation dataset are 217, 28, and 27 videos respectively.

3.3.3 Training

The training process of the baseline model is straightforward. The features of the training dataset from the previous section were fed to the network and computed error with the label by using binary cross-entropy loss function as demonstrated in Equation 7. The label y_n is 0 or 1 and the network output of sigmoid output layer x_n is between 0 and 1.

$$l_n = -[y_n \cdot \ln x_n + (1 - y_n) \cdot \ln(1 - x_n)] \quad (7)$$

We set a batch size equal to 1 video instance per batch. The starting learning rate was 0.00001 which was scheduled to be decreased by 90% for every 5 consecutive epochs that validation loss is not improved. Adaptive moment estimation was used as an optimization algorithm with betas parameters at 0.9 and 0.999. We also set the early stopping criteria that the training will be stopped if 10 consecutive epochs can not improve the best validation loss. The maximum epoch that the network will be trained was also set at 100 epochs. The average loss and

accuracy from the training and validation dataset on each epoch during the training process is shown in Figure 13.

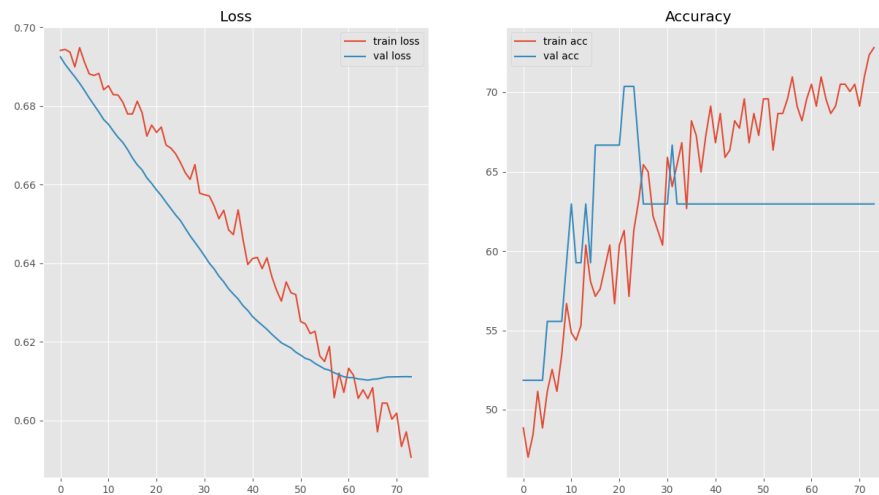


Figure 13 Loss and Accuracy of the training (orange) and validation (blue) dataset of training the baseline model.

3.4 Spatial-Temporal Network model

In this study, we proposed a spatial-temporal network for confusion detection. The video instance is fed to the ResNet-18 model to extract the spatial feature of each frame by convolution and pooling operation. After that, the spatial feature of each frame is passed to the LSTM layer. The LSTM layer learns the temporal information through the sequence of input information by passing the cell state and hidden state from each cell to the next cell that corresponds to the next time step. Then, the hidden state of every cell is fed to a dropout layer, a fully connected layer with a ReLU activation function, another dropout layer, and the output layer of the model which is a fully connected layer with only 1 node. The training process and structure details of this network are described in detail in the following section.

3.4.1 Preprocessing

From the BAUM-1s image-level dataset, we stacked every other frame from the first 2 seconds of each video on the BAUM-1s frame-level dataset to be 1 video instance of the video-level dataset. The video instances that shorter than 2 seconds or have any missing frame from the failure of facial detection were eliminated, results in 875 video instances divided into 214 (24.5%)

confusion instances and 661 (75.5%) non-confusion instances in total. Each instance has the probability to be randomly transformed by one of eight image augmentation techniques including random rotation between -15 to +15 degrees, left to right flipping, contrast adjustment, brightness adjustment, saturate adjustment, salt & pepper noise, gaussian blur, and Gaussian noise on the training process. The examples of the images that were transformed by each technique are shown in Figure 14.

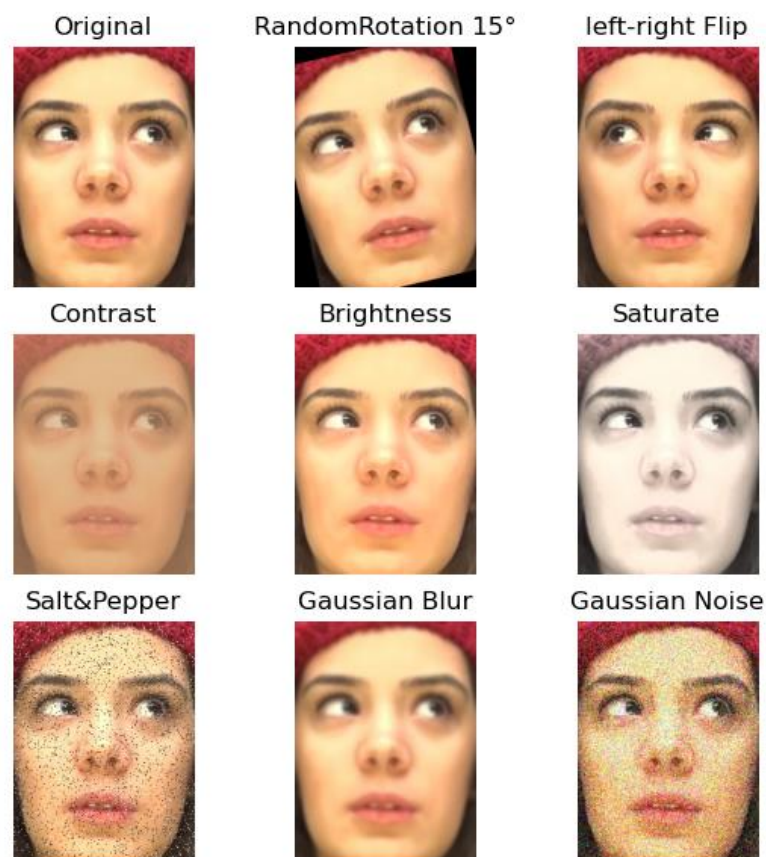


Figure 14 Sample images after applying 8 image augmentation techniques compare to the original facial image.

After the augmentation step, every instance is transformed by other standard steps in order to comply with Pytorch's pretrained model input expectation including be rescaled to be 224 x 224 pixels and normalized the pixel value of the RGB channels by subtracting with 0.485, 0.456, 0.406 then divided by 0.299, 0.224, 0.225 on each channel respectively.

3.4.2 Training Spatial-Temporal network and hyperparameters tuning

From the literature review, we believe that combining spatial and temporal information will increase the performance of confusion detection, but the challenging part is to train the model that proves our hypothesis. Instead of training the whole cascaded model at one time, we decide to train only ResNet-18 as an image classification model for confusion detection first. We chose only the training and validation part of the video-level dataset to train this image classification model. Both training and validation video datasets are converted back to be frame level to perform transfer learning with PyTorch's pretrained ResNet-18 model without freezing any weight parameters. The output fully connected layer of the ResNet-18 is changed from 1,000 nodes to be only 1 node. The data is gradually fed to the model as a batch of 256 images. Binary cross-entropy with logit loss function is used to calculate gradient for backpropagating along with adaptive moment estimation optimization algorithm. The learning rate starts at 0.001 and is scheduled to decrease 90% for every 5 consecutive epochs that fail to improve the validation loss. The maximum epoch is 100 but the early stopping criteria will stop the training after 10 consecutive epochs fail to improve the validation loss.

The spatial-temporal video classification model is the ResNet-18 model which loaded the weight parameters from the trained ResNet-18 image classification model, then connected with LSTM layer, 25% dropout layer, fully connected layer with 64 nodes with ReLU activation function, 30% drop out layer, and fully connected layer with 1 node, respectively. the training process starts from splitting the video-level dataset into training, validation, and testing dataset. The ResNet-18 image classification model is trained by using the training, and validation video-level dataset which converts back to be frame-level datasets by extracting each frame on the video instance to be 1 image instance. After the image classification model is trained, the weight parameters are loaded to another ResNet-18 model on the spatial-temporal video classification model to perform transfer learning on the video-level dataset. the training and validation video-level dataset are used again to train the video model. finally, the testing video-level dataset is used to test the performance of the final model.

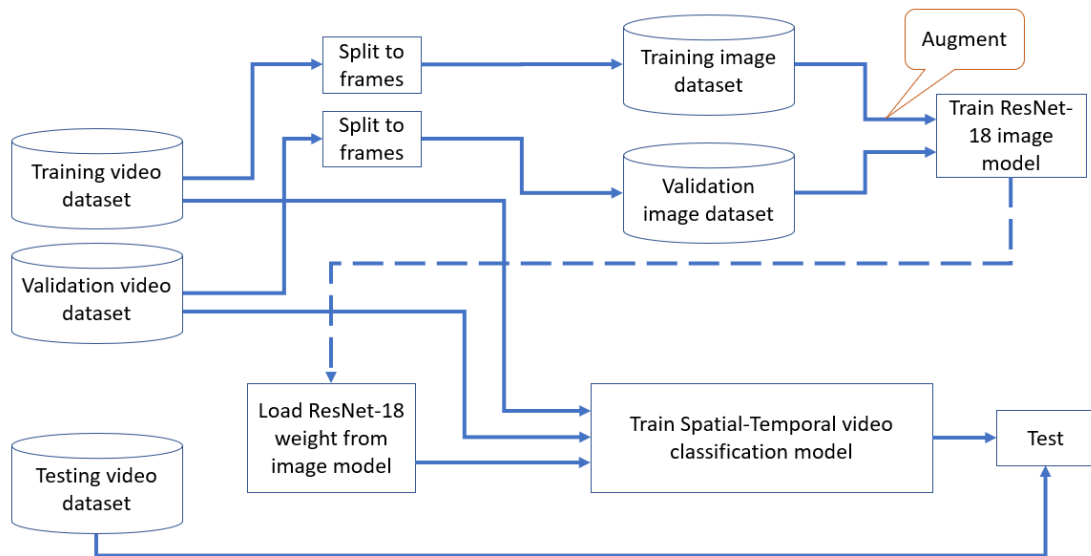


Figure 15 Initial training process of the proposed spatial-temporal confusion detection network.

For the first attempt, we initially chose all 875 video instances from preprocessing step to be divided into 560 training (64%), 140 validation (16%), and 175 testing (20%) instances. The learning rate of the video model starts at 0.001 and is scheduled to decrease 90% for every 5 consecutive epochs that do not improve the validation loss, and the training is stopped after 10 consecutive epochs without validation loss improvement. The training is set to be 100 epochs as maximum. We also use the binary cross-entropy with logit loss function and adaptive moment estimation optimization technique for training this model. The image augmentation is only applied on the ResNet-18 image classification model training with 40% of transformation probability. After loading weight parameters from the image model, the output layer of the ResNet-18 on the video model is modified to be the fully connected layer with 512 nodes without activation function. The training process is demonstrated in Figure 15. The training loss of the ResNet-18 image classification model of this first attempt is smoothly improved, but the validation loss is not, as shown in Figure 16. And the video classification model which loads the weight parameters from the image classification model to be initial weight for transfer learning fails to decrease both training and validation loss except the first to the second epoch. We realized that the image classification model is facing an overfitting problem. Moreover, the testing result of the video model showed that all testing instances were predicted to be a non-confusion class

which makes the model reaches an accuracy that equal to the non-confusion proportion on the dataset at 75%.

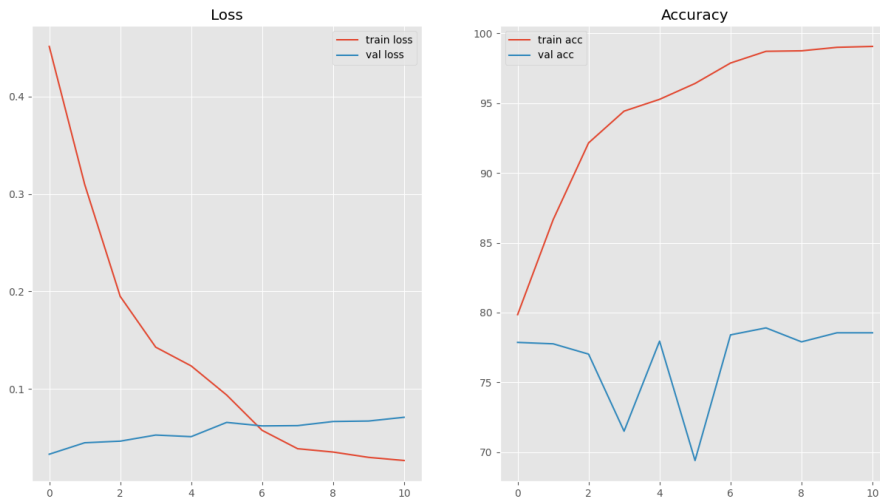


Figure 16 Loss and accuracy on training and validation datasets of ResNet-18 image classification model from the initial training process.

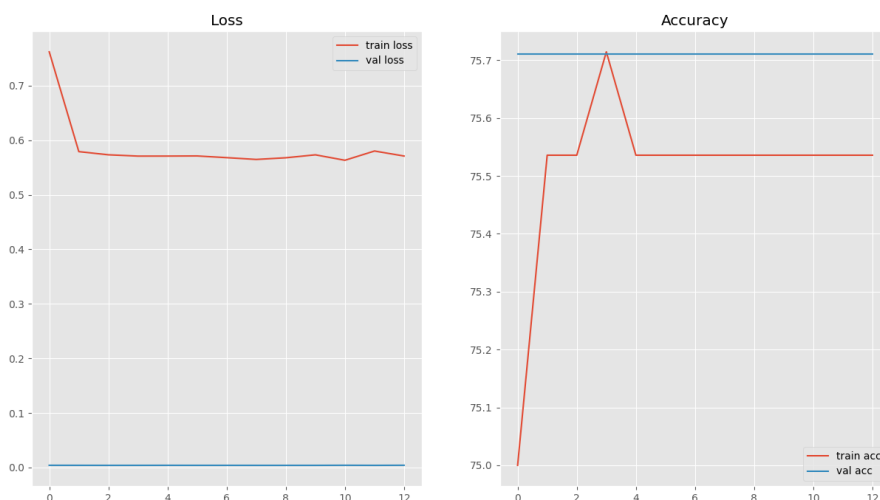


Figure 17 Loss and accuracy on training and validation dataset of spatial-temporal video classification model from the initial training process.

From the problems that we faced on the first attempt, various parameters were adjusted to solve them. For the overfitting problem of the image model, we successfully solved it by merging the training and validation image-level dataset together before randomly re-split to be training, and validation dataset again. This method decreased the difference of training and validation

image-level dataset from the first attempt that each dataset contains only image instances from a different video, hence avoiding overfitting is easier for the image model. The loss and accuracy of the image model after the dataset is adjusted are shown in Figure 19.

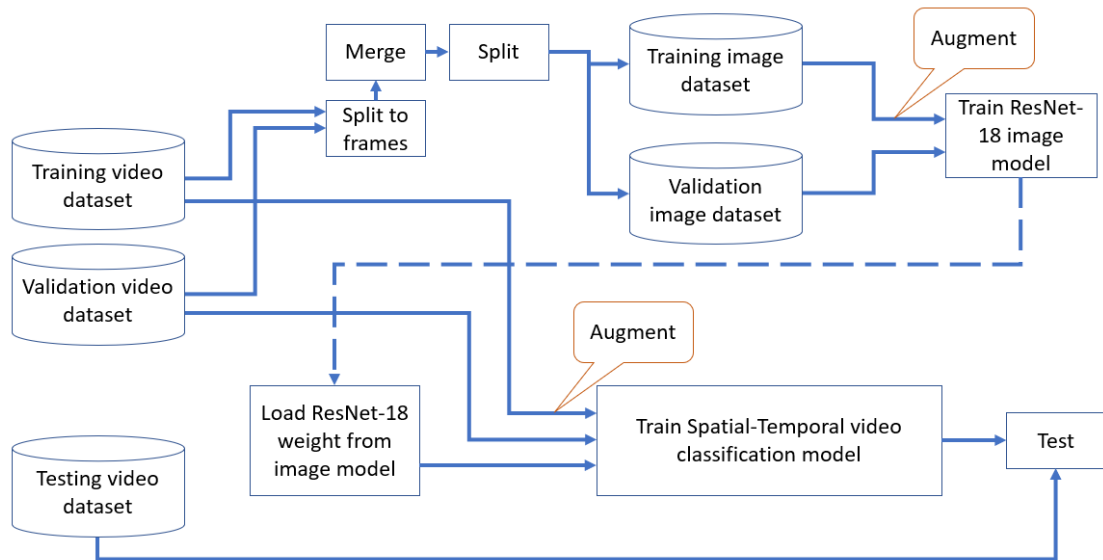


Figure 18 The final training process of the proposed spatial-temporal video classification network.

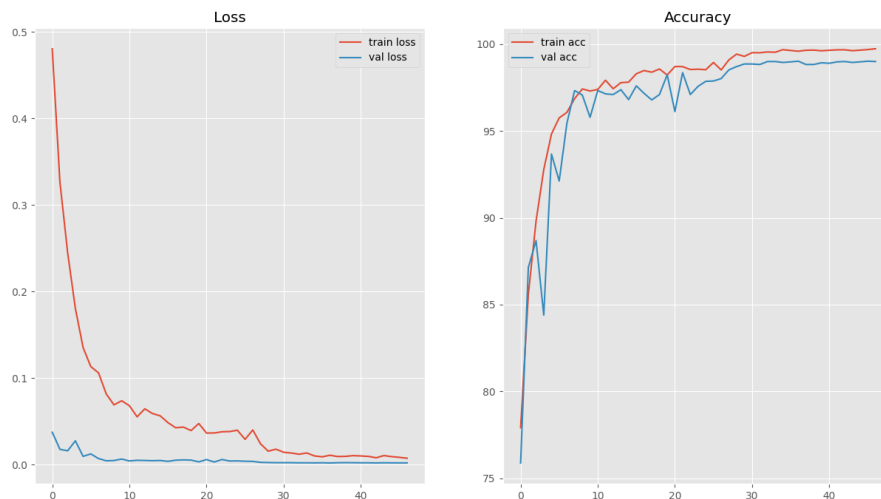


Figure 19 Loss and accuracy of training ResNet-18 image classification model with the image-level dataset from merging the training and validation video-level datasets.

The next problem is that the video model predicts every testing instance as a non-confusion class. Because non-confusion is 75.5% of the initial dataset, we suspect that the model

learns to predict every instance as a non-confusion because it is the fastest way to decrease the training loss. Therefore, we tried to balance the dataset by downsampling the non-confusion video instances to be the same number as confusion instances, which resulted in 428 video instances divide into 214 confusion instances and others are non-confusion instances. But after balancing the dataset, every testing instance was still predicted as non-confusion. However, we also observed another issue from Figure 17 that training loss improves only from the first to the second epoch before keeping the same level until the early stopping criteria is met. This indicates that the model can not learn from the training data properly. Thus, we try to increase the model complexity hoping that will make the model learn the pattern from the data better by adding another layer on the LSTM layer along with changing the multilayer LSTM to be bidirectional. Unfortunately, this adjustment still gives the same testing result. Another cause that makes the model fails to learn from the data is the improper learning rate. In this case, it seems that the learning rate was too high which made the training loss improve only once before stuck to some local minima. Therefore, we decreased the starting learning rate to 0.00001 from the first trial at 0.001. These changes greatly impacted the video model performance to the point that training and validation accuracy reach 100% while the testing accuracy is only 72% which is a strong sign of overfitting. Therefore, in order to increase the model generalization, we adjust various parameters such as increasing augmentation probability, decrease LSTM down to be a single layer, increase dropout probability, etc.

After hyperparameters tuning, we change the ratio of splitting the training, validation, and testing video-level dataset to 80%, 10%, and 10% respectively in order to compare with the baseline model. the final spatial-temporal video classification model is demonstrated in Figure 20. the model structure is the ResNet-18 which the output layer is cut off and then connected to a single LSTM layer without bidirectional, the 50% dropout layer, the fully connected layer with 64 nodes with ReLU activation function, the second 50% dropout layer, and the output fully connected layer with 1 node, respectively. The augmentation is applied on both image and video model training with the probability that each instance will be transformed by one of eight image augmentation techniques at 80%. The training and validation loss and accuracy of the final model are shown in Figure 21.

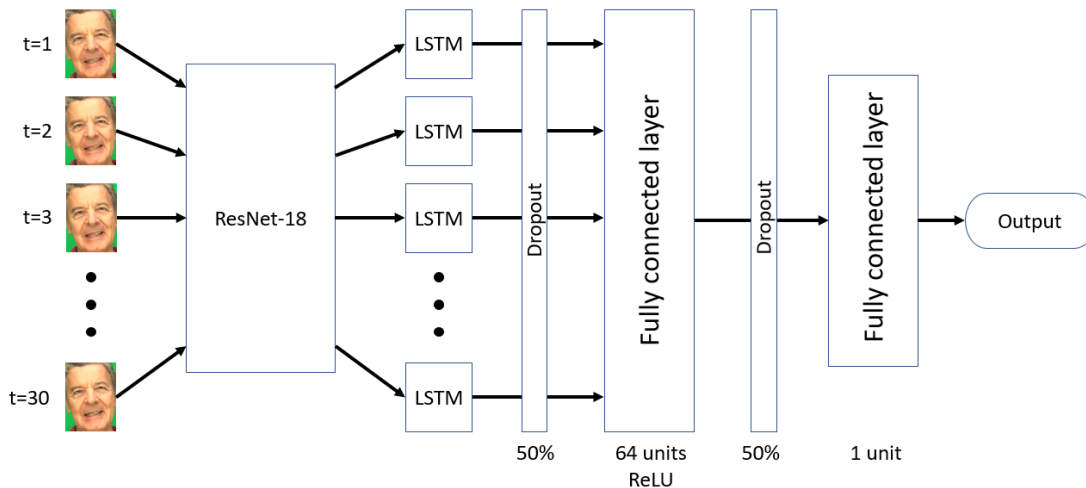


Figure 20 Structure of the proposed spatial-temporal network for confusion detection.

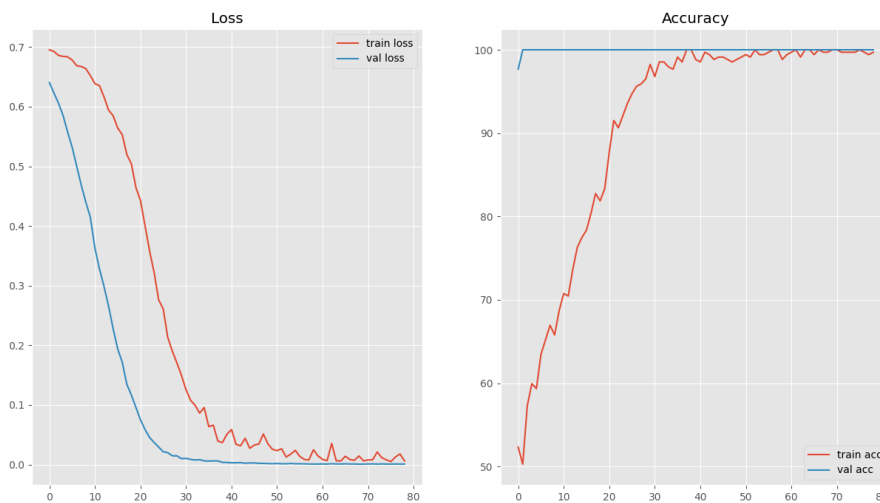


Figure 21 The loss and accuracy from training the final spatial-temporal confusion detection network.

CHAPTER 4 TESTING PROCEDURE AND EXPERIMENTAL RESULT

4.1 Performance Measurements

The performance of the model was measured using binary classification metrics. the prediction result can be summarized as a confusion matrix that divides into 4 scenarios including true positive, true negative, false positive, and false negative.

		Prediction	
		Negative	Positive
Ground truth	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Figure 22 Confusion Matrix

From the confusion matrix, we can calculate 4 measurements which interpret the different aspect of the model performance. The first measurement is accuracy which can be calculated by Equation 8. This measurement shows the probability that the classifier predicts the class of samples correctly.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (8)$$

When the classifier predicts a sample to be positive, the probability that the sample is really positive can be measure by precision. And for all positive samples, the number of a sample that the classifier correctly classifies samples as positive is recall.

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

A classifier might have to trade-off between precision and recall. In the extreme case, the classifier that predicts every sample to be positive will have 100% of recall because the false negative is 0 while the precision is low from the high amount of false positive. On the other hand, if the classifier classifies a very little number of samples to be positive, it might reach very high

precision along with low recall. This problem leads to an F1-score which combines precision and recall into a single measurement.

$$F1 - score = 2 \cdot \frac{precision \times recall}{precision + recall} \quad (11)$$

Another measurement to compare between classifiers is the Receiver Operating Characteristic (ROC) curve which is a graph of True Positive Rate (TPR) against False Positive Rate (FPR) on each classification threshold. Most binary classifiers predict the probability of a sample before distinguishing samples with a higher probability than the threshold to one class, and others to another class. When the classifying threshold is changed, the predicted class of a sample is also changed. If the threshold to classify a sample to be positive class is high, the TPR and FPR are usually high together. Therefore, the classifier that has high TPR while maintaining low FPR is preferred. Performance of classifiers can be compared with Area Under the Curve (AUC) which 1 is ideal and a classifier with more AUC is better than another with lower.

$$TPR = \frac{TP}{FN + TP} \quad (12)$$

$$FPR = \frac{FP}{TN + FP} \quad (13)$$

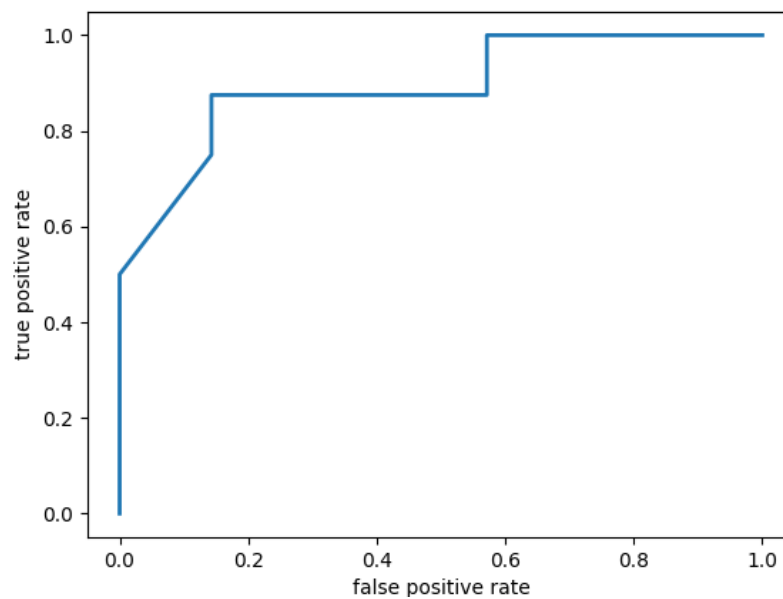


Figure 23 ROC curve

4.2 Comparative evaluation using BAUM-1 database

In this section, we present the testing result of our proposed model compared to the baseline model on BAUM-1s and BAUM-1a video datasets. The BAUM-1s testing video dataset of the baseline model has 28 instances divided into 14 confusion instances and 14 non-confusion instances while the proposed model BAUM-1s testing dataset, which does not lose data from the failure of AUs activation intensity extraction, have 43 instances divided into 21 confusion and 22 non-confusion instances. According to the significant difference in the number of testing instances, we decided to use the BAUM-1a dataset as a secondary testing dataset which we selected only video instances that have the AUs activation intensity features on every other frame on the first 2-second duration, resulted in 124 video instances divided into 15 confusion and 109 non-confusion. We consider testing both models on the same dataset as a fair comparison, even though the proportion of confusion instances is very low, but it reflects the frequency of confusion expression that much less than non-confusion in daily life.

We compared the performance of the two models in 5 measurements including accuracy, precision, recall, F1-score, and AUC. On the BAUM-1s testing dataset, the baseline model achieves the measurements at 68%, 65%, 79%, 71% and 69%, while the proposed method is at 74%, 75%, 71%, 73% and 74% respectively. The result shows that our proposed method outperforms the baseline model on almost every measurement except recall.

Result from testing on BAUM-1s testing dataset

Baseline model				Proposed model					
Actual	Predicted	non-confusion	confusion	total	Actual	Predicted	non-confusion	confusion	total
		non-confusion	confusion				non-confusion	confusion	
non-confusion		8	6	14	non-confusion		17	5	22
confusion		3	11	14	confusion		6	15	21
total		11	17	28	total		23	20	43

Figure 24 Confusion matrix of the baseline and proposed model on their BAUM-1s testing dataset.

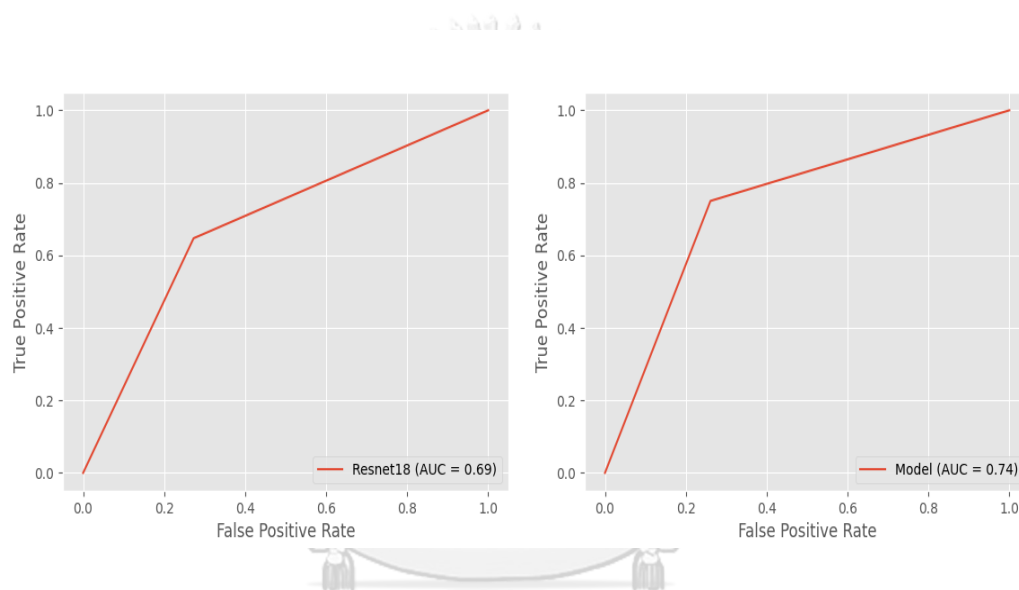


Figure 25 ROC curve of the baseline model (left) and proposed model (right) on BAUM-1s testing dataset.

We also tested the models on BAUM-1a which showed a different result from the BAUM-1s dataset, the performance of baseline model which has accuracy, precision, recall, F1-score, and area under ROC at 67%, 19%, 53%, 28%, and 55% is almost higher than the proposed model at 73%, 15%, 27%, 20%, and 52% respectively. All testing results are summarized in Table 6.

Result from testing on BAUM-1a testing dataset

Baseline model				Proposed model					
Actual	Predicted	non-confusion	confusion	total	Actual	Predicted	non-confusion	confusion	total
		non-confusion	confusion				non-confusion	confusion	
non-confusion		75	34	109	non-confusion		87	22	109
confusion		7	8	15	confusion		11	4	15
total		82	42	124	total		98	26	124

Figure 26 Confusion matrix of the baseline and proposed model on the BAUM-1a dataset.

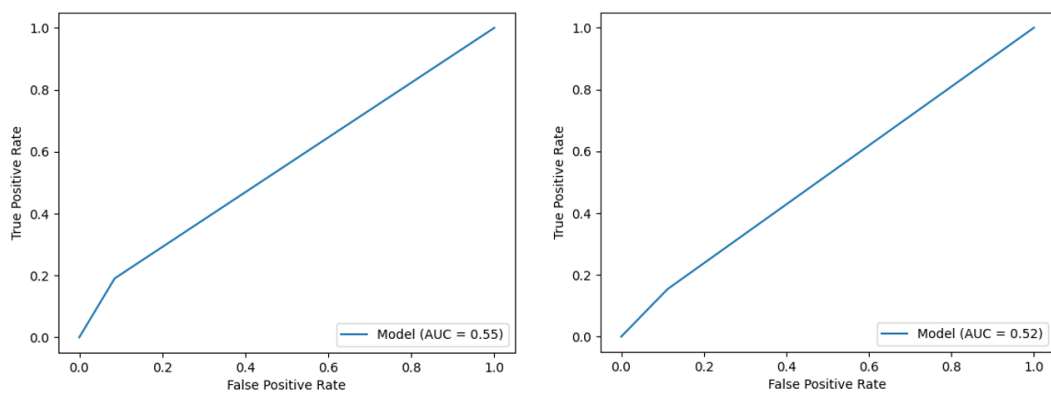


Figure 27 ROC curve of the baseline model (left) and the proposed model (right) on the BAUM-1a dataset.

Table 6 Testing results from the baseline and the proposed model on every measurement.

Measurement	Baseline model		Proposed model	
	BAUM-1s	BAUM-1a	BAUM-1s	BAUM-1a
accuracy	68%	67%	74%	73%
precision	65%	19%	75%	15%
recall	79%	53%	71%	27%
F1-score	71%	28%	73%	20%
AUC	69%	55%	74%	52%

We also analyzed the confusion matrix of ground truth emotion/mental states from the original BAUM-1 annotation and prediction results. The confusion class does not show any additional insight, because, in the BAUM-1a video dataset, Unsure is the only member of the

confusion group. Whereas on the non-confusion, both baseline and proposed model show pattern of FPR as shown in Figures 28, and 29.

Table 7 Confusion Matrix of the baseline model on BAUM-1a dataset

Actual	Predicted		
	Non-Confusion	Confusion	Total
Non-Confusion	75	34	109
Anger	9	10	19
Boredom	8	4	12
Disgust	14	1	15
Fear	9	7	16
Happiness	11	2	13
Interest	12	1	13
Sadness	12	9	21
Confusion	7	8	15
Unsure	7	8	15
Total	82	42	124

Table 8 Confusion Matrix of the proposed model on BAUM-1a dataset

Actual	Predicted		
	Non-Confusion	Confusion	Total
Non-Confusion	87	22	109
Anger	15	4	19
Boredom	9	3	12
Disgust	13	2	15
Fear	12	4	16
Happiness	13		13
Interest	12	1	13
Sadness	13	8	21
Confusion	11	4	15
Unsure	11	4	15
Total	98	26	124

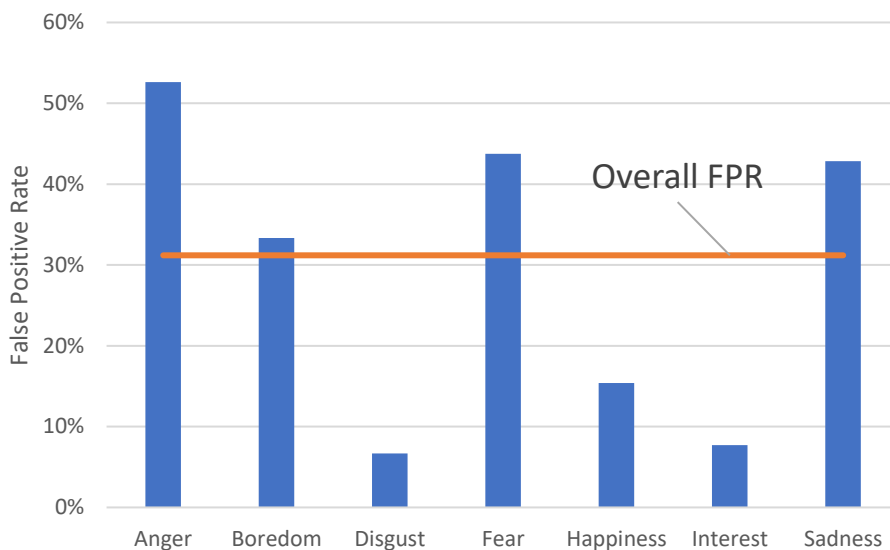


Figure 28 FPR of the baseline model on BAUM-1a testing dataset.

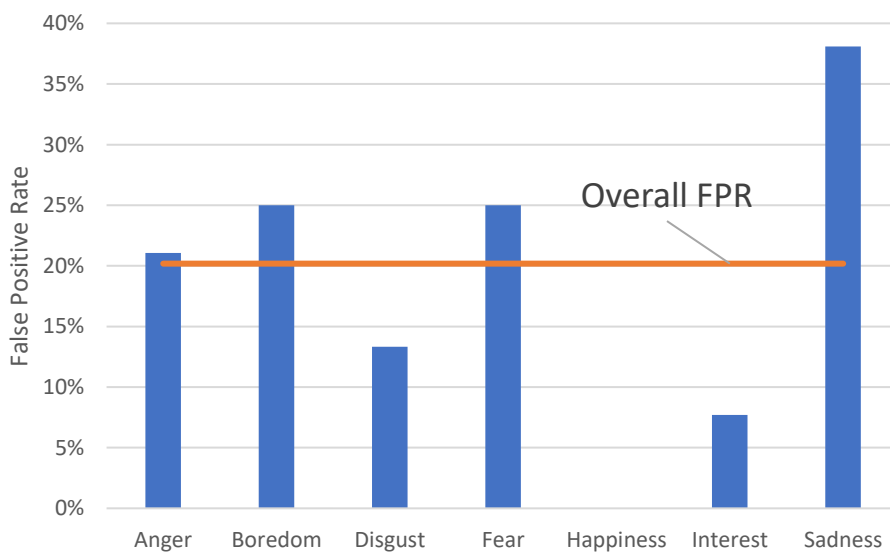


Figure 29 FPR of the proposed model on BAUM-1a testing dataset.

The anger, boredom, fear, and sadness have higher FPR than the overall FPR on both baseline and proposed model while disgust, happiness, and interest are lower. This pattern indicates that those emotion/mental states with high FPR have similarity with confusion more than others with low FPR, thus more difficult for the model to correctly classify them.

4.3 Comparative evaluation using our CUPIC-Confusion video dataset

In addition to testing the proposed model on the BAUM-1 database, we performed the testing on our confusion video dataset which is recorded from an uncontrolled environment. we believe that testing on the realistic data will result in a more realistic performance of the confusion detection algorithm.

4.3.1 Testing procedure

Testing methodology on the confusion video dataset is different from the BAUM-1 dataset because we labeled the confusion interval on each video instead of label each video with one emotion/mental state. We compute the model prediction by feeding one by one of every other frame from the video to the facial detector. The facial image from each frame is sequentially preprocessed and appended to the instance array. When the instance array stores 30 frames of the facial images, the instance array is fed to the spatial-temporal network to predict the probability of confusion. The predicted class was applied to every frame represented by the instance array on the video. After that, the instance array is reset to be empty for storing a facial image of the next frame from the video. In case the facial detector fails to detect the face region on any frame, every frame that corresponds to the first to the last on the instance array is predicted as null automatically and the instance array is also reset. Thus, the instance array that fed to the network always contains 30 facial images from consecutive every other frame of the video. The instance that got above 50% predicted probability is classified as confusion on the final prediction.

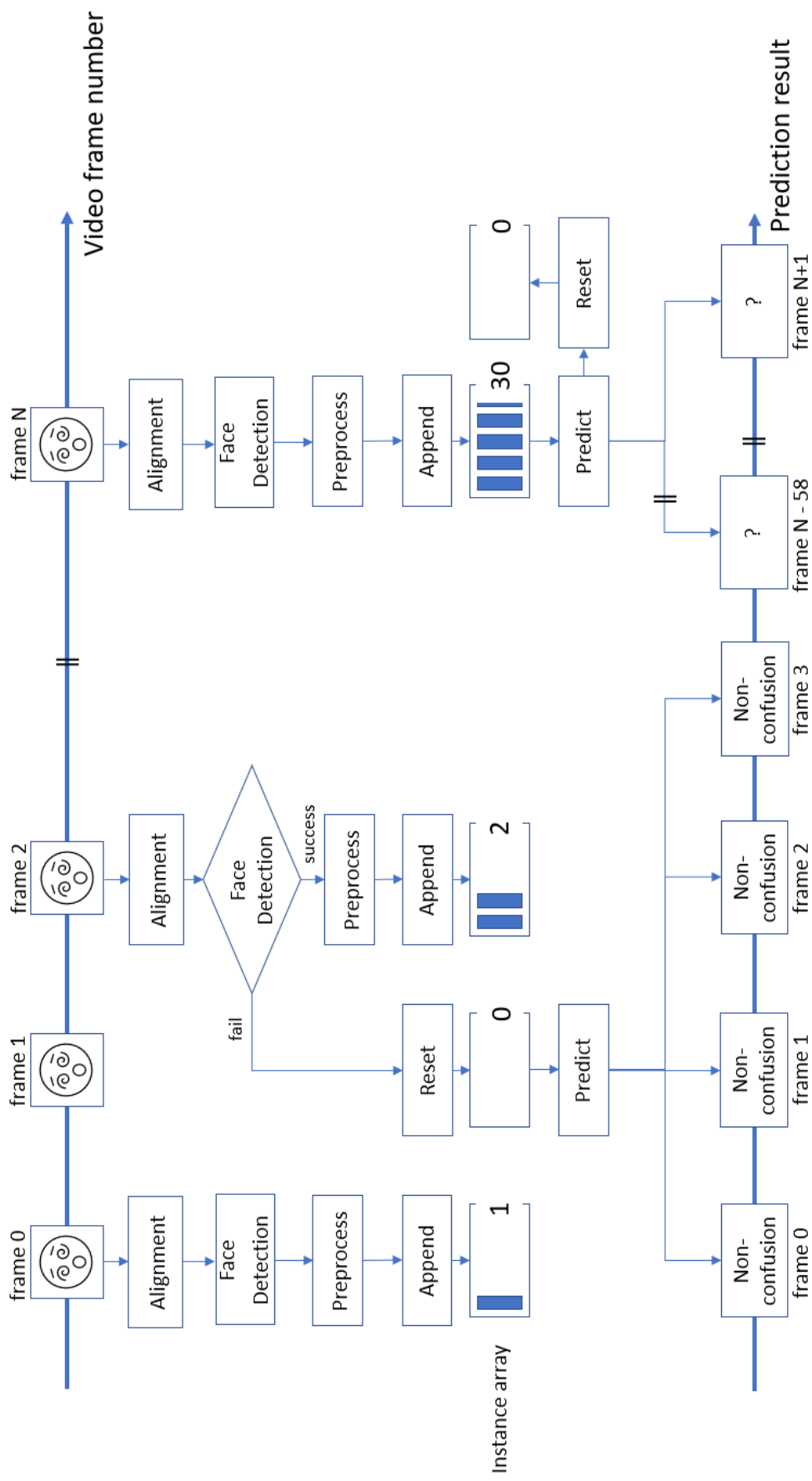


Figure 30 Testing procedure on the confusion video dataset.

4.3.2 Testing result

The testing result of each video is recorded as each scenario on the confusion matrix, including true negative, false positive, false negative, and true positive. We summarize the number of frames as a percentage of 764,284 frames for convenience of interpretation as shown in Figure 31 and others measurements that are calculated from the confusion matrix are shown in Table 9.

		Predicted		total
		non-confusion	confusion	
Actual	non-confusion	65.15%	31.71%	96.86%
	confusion	1.97%	1.17%	3.14%
	total	67.12%	32.88%	100.00%

Figure 31 Confusion matrix of the testing result on the confusion video dataset.

Table 9 Performance of the proposed model on the confusion video dataset

Measurement	
accuracy	66%
precision	4%
recall	37%
F1-score	7%

Not only accuracy, but we also measure the speed of each step on the testing steps. From averaging the average time of each video, rotating the frame to be perpendicular, and detecting facial region take 0.29 seconds per frame. Preprocessing which includes cropping the facial region, resizing the facial image, converting the facial image from NumPy array to tensor, and normalizing takes another 0.014 seconds. The spatial-temporal network inference time which waits until the facial image from 30 consecutive every other frame is collected to 30 frames is 0.04 seconds per instance, and the time for labeling every frame on the input frame range with the prediction result is 0.003 seconds. For 60 frames on the testing video, the testing procedure performs 30 times of face detection, 30 times of preprocessing, 1 time of inferencing, and 1 time

of labeling which takes 9.16 seconds. The testing speed can be converted to 6.55 fps which too low for real-time implementation.



CHAPTER 5 CONCLUSION AND FUTURE WORKS

5.1 Conclusion

For confusion detection from FER, the spatial-temporal network is proposed. The network learns the spatial information and temporal information from the input data by CNN and LSTM structures respectively. The proposed method clearly outperforms the baseline LSTM network, however, under the uncontrolled environment, the accuracy is significantly dropped which we believe it is caused by the variety of environments such as lighting conditions, camera angles, and video qualities. Besides, the inference speed of the proposed model is 0.04 seconds per 30 frames while facial detection which requires 0.29 seconds per frame is the bottleneck of the whole process. However, we discovered that automatic AU activation intensity extraction is a challenging task which makes the method of the baseline model harder to be implemented compared to the proposed method that extracts the spatial feature of the input image by using CNN. Besides, from analyzing the testing result on the BAUM-1a dataset, both baseline and proposed model show the result in the same way that Disgust, Happiness, and Interest have lower FPR than overall which indicates ease of differentiation from confusion. We investigated this topic on BAUM-1s frame-level dataset by calculating Pearson correlation coefficient similarity of average AUs activation intensity of each emotion/mental state with confusion as shown in Figure 32. The emotion/mental states that have lower similarity than overall non-confusion are Disgust, Happiness, Surprise, and Contempt. The similarity validates the model's testing result that each emotion has a different similarity with confusion thus, different levels of difficulty to be differentiated from confusion.

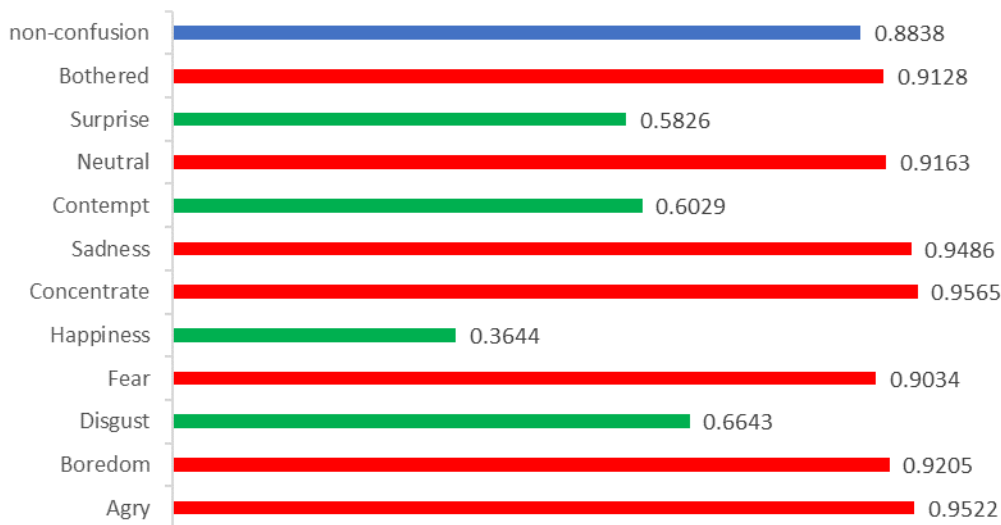


Figure 32 Pearson correlation coefficient similarity of each emotion/mental states with confusion, computed from average activation intensity of 17 AUs.

5.2 Recommendations for future works

Confusion video dataset especially that recorded under uncontrolled environment is still hard to find. Data collection should be performed with a reliable method of annotation such as majority vote from annotators.

Using other information along with facial expression is also a potential approach to improve the confusion detector performance. Cai et al [29] already studied using both audio and visual information for emotions classification (Angry, Excite, Neutral, and Sad) on the IEMOCAP dataset [30]. Their study reveals that a multimodal model got better performance when compared to a unimodal model. Another interesting piece of information for confusion detection is from the user interface field, the mouse and eye movements dataset [31] had collected from the interaction of participants while they were performing instructed experimental tasks on a web application. But the information that relates to subject interaction such as conversation or mouse movement does not often happen during large-scale communication in which the audience only perceives information. therefore, we believe that information about the content that the subject perceiving is more suitable feature.

Even though the proposed method which uses CNN to extracts spatial features from the image reaches higher accuracy than the baseline model that extracts AUs from the image, the

Openface 2.0 developer states on their GitHub page that AU extraction from a video file gets a better result than extracting from still an image. Hence, a better AU extraction might lead to better confusion detection.

For the temporal features, we select to study only the first 2-second duration from each video which the expression might not fully present, or the most intense expression might occur at other intervals of the video. Moreover, the length of confusion expression might not be 2 seconds. The evidence that supports this hypothesis is that the average length of confusion video on the BAUM-1 database is 6 seconds which close to the average confusion interval in our confusion video dataset at 6.5 seconds. The reason that we did not choose to study 6 seconds per instance because the number of a clip that long enough is low (277 from 1134 on BAUM-1s) which we concern about model generalization from training with a small dataset. Therefore, we believe that adding an algorithm to automatically focus on the important interval of the video such as selecting peak frame in [1], and adaptive key frame interval in [28] is an interesting topic to study further.

REFERENCES

1. Zhalehpour, S., et al., *BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States*. IEEE Transactions on Affective Computing, 2017. **8**(3): p. 300-313.
2. Borges, N., et al. *Classifying Confusion: Autodetection of Communicative Misunderstandings using Facial Action Units*. in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. 2019.
3. Ekman, P. and W.V. Friesen, *Constants across cultures in the face and emotion*. Journal of Personality and Social Psychology, 1971. **17**(2): p. 124-129.
4. Viola, P. and M. Jones. *Rapid object detection using a boosted cascade of simple features*. in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. 2001.
5. Sagonas, C., et al. *300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge*. in *2013 IEEE International Conference on Computer Vision Workshops*. 2013.
6. *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS), 2nd ed.* What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS), 2nd ed., ed. P. Ekman and E.L. Rosenberg. 2005, New York, NY, US: Oxford University Press. xxi, 639-xxi, 639.
7. Berlyne, D.E., *Conflict, arousal, and curiosity*. Conflict, arousal, and curiosity. 1960, New York, NY, US: McGraw-Hill Book Company. xii, 350-xii, 350.
8. Keltner, D. and M.N. Shiota, *New displays and new emotions: A commentary on Rozin and Cohen (2003)*. Emotion, 2003. **3**(1): p. 86-91.
9. Ellsworth, P.C., *Confusion, concentration, and other emotions of interest: Commentary on Rozin and Cohen (2003)*. Emotion, 2003. **3**(1): p. 81-85.
10. Smith, C.A. and P.C. Ellsworth, *Patterns of cognitive appraisal in emotion*. Journal of Personality and Social Psychology, 1985. **48**(4): p. 813-838.
11. Silvia, P.J., *Confusion and interest: The role of knowledge emotions in aesthetic experience*. Psychology of Aesthetics, Creativity, and the Arts, 2010. **4**(2): p. 75-80.

12. Keltner, D. and J. Haidt, *Approaching awe, a moral, spiritual, and aesthetic emotion*. *Cogn Emot*, 2003. **17**(2): p. 297-314.
13. Konečňni, V., *The aesthetic trinity: Awe, being moved, thrills*. *Bulletin of Psychology and the Arts*, 2005. **5**: p. 27-44.
14. Ludden, G.D.S., H.N.J. Schifferstein, and P. Hekkert, *Visual–Tactual Incongruities in Products as Sources of Surprise*. *Empirical Studies of the Arts*, 2008. **27**(1): p. 61-87.
15. Darwin, C., *The expression of the emotions in man and animals*. The expression of the emotions in man and animals. 1872, London, England: John Murray. vi, 374-vi, 374.
16. Rozin, P. and A.B. Cohen, *High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans*. *Emotion*, 2003. **3**(1): p. 68-75.
17. Durso, F.T., K.M. Geldbach, and P. Corballis, *Detecting Confusion Using Facial Electromyography*. *Human Factors*, 2012. **54**(1): p. 60-69.
18. Cordaro, D.T., *Universals and Cultural Variation in Emotional Expression*. UC Berkeley Electronic Theses and Dissertations, 2014.
19. He, K., et al., *Deep Residual Learning for Image Recognition*. arXiv e-prints, 2015: p. arXiv:1512.03385.
20. Erwianda, M.S.F., et al. *Improving Confusion-State Classifier Model Using XGBoost and Tree-Structured Parzen Estimator*. in *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. 2019.
21. J. Gordon Betts, K.A.Y., James A. Wise, Eddie Johnson, Brandon Poe, Dean H. Kruse, Oksana Korol, Jody E. Johnson, Mark Womble, Peter DeSaix, *Anatomy and Physiology*. 2013, OpenStax.
22. Jung, H., et al. *Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition*. in *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015.
23. Zhang, K., et al., *Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks*. *IEEE Transactions on Image Processing*, 2017. **26**(9): p. 4193-4203.
24. Li, S. and W. Deng, *Deep Facial Expression Recognition: A Survey*. arXiv e-prints, 2018: p. arXiv:1804.08348.
25. Shi, Z., et al. *Automatic Academic Confusion Recognition In Online Learning Based On*

- Facial Expressions*, in *2019 14th International Conference on Computer Science & Education (ICCSE)*. 2019.
26. Simonyan, K. and A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv e-prints, 2014: p. arXiv:1409.1556.
 27. Zhou, X., D. Wang, and P. Krähenbühl, *Objects as Points*. arXiv e-prints, 2019: p. arXiv:1904.07850.
 28. Baltrusaitis, T., et al. *OpenFace 2.0: Facial Behavior Analysis Toolkit*. in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. 2018.
 29. Cai, L., J. Dong, and M. Wei. *Multi-Modal Emotion Recognition From Speech and Facial Expression Based on Deep Learning*. in *2020 Chinese Automation Congress (CAC)*. 2020.
 30. Busso, C., et al., *IEMOCAP: interactive emotional dyadic motion capture database*. *Language Resources and Evaluation*, 2008. **42**(4): p. 335.
 31. Hucko, M., R. Moro, and M. Bielikova, *Confusion Detection Dataset of Mouse and Eye Movements*, in *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 2020, Association for Computing Machinery: Genoa, Italy. p. 281–286.



จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

VITA

NAME Nun Vanichkul

DATE OF BIRTH 27 June 1995

PLACE OF BIRTH Phramonkud Hospital, Bangkok

INSTITUTIONS ATTENDED Bachelor of Engineering, Industrial Engineering, Chulalongkorn University 2013
Master of Science, Computer Science, Chulalongkorn University 2019

HOME ADDRESS 49 Punnavithi 17, Bangchak, Phra Khanong, Bangkok 10260



จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY