การจับความรู้สึกของคนจากใบหน้าด้วยเทคนิคปัญญาประดิษฐ์

Classification of human facial expression using artificial intelligence techniques

ศศิชา บุญเก่า[1], วิทิต ฉัตรรัตนกุลชัย[2*]

Sasicha Boonkao[1], Withit Chatlatanagulchai[2*]

Received: 18 July 2019; Revised: 19 August 2019; Accepted: 10 September 2019

บทคัดย่อ

ในปัจจุบันมีการนำเครื่องจักรไปใช้ในการทำงานที่แตกต่างกันเพิ่มขึ้นเรื่อย ๆในสังคม จึงนำไปสู่ความหลากหลายของงานรวมถึงสิ่งที่ซับซ้อนมาก ดังนั้นการรับรู้ของเครื่องจักรจึงต้องการให้เครื่องเข้าใจเกี่ยวกับสภาพแวดล้อมและความเข้าใจของคู่สนทนา โดยในบทความนี้ทำการวิจัยเกี่ยวกับการเรียนรู้ของเครื่องจักรที่ตระหนักถึงอารมณ์ที่แสดงออกทางใบหน้าของมนุษย์ โดยการพัฒนาในงานวิจัยนี้ใช้เทคนิคทางปัญญาประดิษฐ์ผสมผสานกันระหว่างการจำแนกอารมณ์และการเรียนรู้เชิงลึกโดยใช้เครือข่ายประสาทเทียมแบบคอนโวลูชันที่ถูกฝึกมาแล้วร่วมกับการเรียนรู้แบบถ่ายโอนเพื่อระบุอารมณ์ความรู้สึกที่สำคัญของมนุษย์ทั้งเจ็ดอารมณ์ ได้แก่ ความโกรธ ความรังเกียจ ความกลัว ความสุข ความเศร้า ความประหลาดใจและความเป็นกลาง โดยเราสามารถใช้ประโยชน์จากชุดเครื่องมือและถ่ายโอนเทคนิคการเรียนรู้เชิงลึกนี้ เช่น การตระหนักถึงอารมณ์ที่แสดงออกทางใบหน้าของผู้สูงอายุที่แสดงออกมาว่าต้องการความช่วยเหลือเมื่อใด โดยผลการทดลองที่ได้งานวิจัยนี้สามารถตรวจจับอารมณ์ของมนุษย์ได้อย่างมีประสิทธิภาพ มีความแม่นยำและสามารถนำไปพัฒนาต่อยอดเป็นผลิตภัณฑ์ในเชิงพาณิชย์ได้

คำสำคัญ: เทคนิคทางปัญญาประดิษฐ์ การจำแนกอารมณ์ เทคนิคการเรียนรู้เชิงลึก

Abstract

Machines are increasingly being used for different work in society for a variety of tasks including very complex things. Therefore, the machine's perception requires that it understands the environment and the interlocutor. This article reports research on machinery that learns to recognize the emotions expressed in human faces. The development of this research uses artificial intelligence techniques, combined with emotional classification and in-depth learning by using an artificial neural network that has been trained together with transfer learning to identify the seven major emotions of human emotions, anger, disgust, fear, happiness, sadness, surprise and neutral. We can take advantage of the tool set and transfer this deep learning technique, such as recognizing the emotions expressed in the face of the elderly that show when they need help. The results of the experiment, show that machines can effectively detect human emotions with precision and can be used to develop into commercial products.

Keywords: Artificial intelligence techniques, Emotional classification, deep learning techniques

---

[1] นิสิต, ภาควิชาวิศวกรรมเครื่องกล คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ เลขที่ 50 ถนนงามวงศ์วาน แขวงลาดยาว เขตจตุจักร กรุงเทพฯ 10900

[2] อาจารย์, ภาควิชาวิศวกรรมเครื่องกล คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ เลขที่ 50 ถนนงามวงศ์วาน แขวงลาดยาว เขตจตุจักร กรุงเทพฯ 10900

** ติดต่อ: fengwtc@ku.ac.th, 0-2797-0999 ต่อ 1858*

[1] Student, Department of Mechanical Engineering, Faculty of Engineering, Kasetsart University, 50 Ngamwongwan Road, Ladyao Sub-district, Chatuchak District, Bangkok 10900

[2] professor,Department of Mechanical Engineering, Faculty of Engineering, Kasetsart University, 50 Ngamwongwan Road, Ladyao Sub-district, Chatuchak District, Bangkok 10900

บทนำ

อารมณ์คือ ความรู้สึกที่เกิดจากการได้รับการกระทบจากสิ่งเร้า อารมณ์มีได้ทั้งทางบวกและทางลบ เป็นได้ทั้งความพึงพอใจและความรู้สึกไม่สมปรารถนา พฤติกรรมของมนุษย์เราเป็นจำนวนมากอยู่ภายใต้การควบคุมของอารมณ์ อารมณ์จึงมีความสำคัญและเป็นเรื่องที่เราจะต้องเรียนรู้และเข้าใจ

ในปัจจุบันมีการนำเครื่องจักรไปใช้ในการทำงานที่แตกต่างกันเพิ่มขึ้นเรื่อยๆในสังคม จึงนำไปสู่ความหลากหลายของงานรวมถึงสิ่งที่ซับซ้อนมาก ดังนั้นการรับรู้ของเครื่องจักรจึงต้องการให้เครื่องเข้าใจเกี่ยวกับสภาพแวดล้อมและความเข้าใจของคู่สนทนา โดยการแสดงออกทางสีหน้าของมนุษย์สามารถบ่งบอกถึงอารมณ์และความรู้สึกในขณะนั้นได้เป็นอย่างดี การรับรู้อารมณ์ของมนุษย์จึงสามารถนำไปวิเคราะห์เพื่อใช้ประโยชน์ได้หลายด้าน เช่น ด้านการพัฒนาหุ่นยนต์เพื่อให้หุ่นยนต์สามารถรับรู้การแสดงอารมณ์ของมนุษย์ได้ ด้านการแพทย์ ช่วยให้แพทย์สามารถรับรู้อารมณ์ของผู้ป่วยเพื่อนำไปวินิจฉัย และเลือกวิธีรักษาได้อย่าแม่นยำและมีประสิทธิภาพมากยิ่งขึ้นและใช้เป็นส่วนหนึ่งของการวิเคราะห์โต้ตอบได้อย่างเหมาะสม จึงทำให้มีการศึกษาค้นคว้าและวิจัยเกี่ยวกับกระบวนการการรับรู้อารมณ์ของมนุษย์กันอย่างแพร่หลาย ซึ่งนำไปสู่การใช้เทคนิคทางปัญญาประดิษฐ์ในการรู้จำอารมณ์ของมนุษย์โดยส่วนใหญ่จะใช้การรู้จำอารมณ์จากการแสดงออกทางสีหน้า ซึ่งมีรูปแบบของอารมณ์พื้นฐานอยู่เจ็ดอารมณ์ด้วยกัน คือ ความโกรธ(anger), ความรังเกียจ(disgust), ความกลัว(fear), ความสุข(happiness), ความเศร้า(sadness), ความประหลาดใจ(surprise) และความเป็นกลาง(neutral)

วัตถุประสงค์และขอบเขต

ออกแบบและพัฒนาระบบให้สามารถเรียนรู้อารมณ์ที่แสดงออกอารมณ์ทางใบหน้าของมนุษย์โดยใช้ประโยชน์จากเทคนิคทางปัญญาประดิษฐ์

ใช้เทคนิคทางปัญญาประดิษฐ์ผสมผสานกันระหว่างการจำแนกอารมณ์และการเรียนรู้เชิงลึกโดยใช้เครือข่ายประสาทเทียมแบบคอนโวลูชันที่ถูกฝึกมาแล้วร่วมกับการเรียนรู้แบบถ่ายโอน เพื่อระบุอารมณ์ความรู้สึกที่สำคัญของมนุษย์ทั้งเจ็ดอารมณ์ ได้แก่ ความโกรธ ความรังเกียจ ความกลัว ความสุข ความเศร้า ความประหลาดใจ และความเป็นกลาง

การออกแบบเครือข่ายประสาทเทียม

1. เครือข่ายประสาทเทียมแบบคอนโวลูชัน (convolutional neural network)

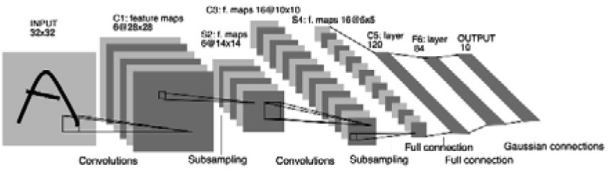

Figure 1 convolutional neural network of LeNet (Sorce: LeCun et al. (1998))

LeNet ประกอบด้วยชั้น (layer) ต่างๆดังนี้

1.1 Input layer

สำหรับใน Figure 1 input data เป็น matrix ขนาด แต่ละ element ของ matrix มีค่า 0-255 (เลข 8 บิท) โดยเลข 0 แทนสีดำ เลข 255 แทนสีขาว และเลขระหว่างกลางแทนสีเทาที่มีความ เข้มแตกต่างกัน (grayscale) หากรูปที่นำมาเป็น input เป็นรูปสี input data จะเป็น matrix ขนาด n x n x 3 เมื่อ n แทนจำนวนจุด (pixel) ของภาพในแต่ละมิติ และ 3 คือความเข้มของแสงในแต่ละ แม่สี red, green, blue (RGB)

1.2 Convolution layer ชั้นที่ 1 (C1)

เป็นชั้นที่เป็นที่มาของคำว่าเครือข่ายประสาทเทียมแบบคอนโวลูชัน รูปด้านล่าง แสดงตัวอย่างการทำคอนโวลูชันของ convolution layer โดย input data เป็น matrix ขนาด 5×5×3 แทนภาพสี ขนาด 5×5 pixels สังเกตว่าในรูปมีการเพิ่มเลข 0 เข้าไปที่ขอบของภาพ โดยมีความหนา 1 pixel เรียกเทคนิคนี้ว่า zero padding เพื่อเป็นการอนุรักษ์ความสำคัญของ pixel ที่อยู่ตรงขอบของภาพ การทำ zero padding ในรูปนี้ ทำให้เกิด input data ขนาด 7×7×3 ในรูปมีค่าถ่วงน้ำหนัก 2 ชุดคือ w_0 และ w_1 ทั้งคู่มีขนาด ค่าถ่วงน้ำหนักนี้มีอีกชื่อว่า filter หรือ kernel เมื่อนำ filter มา convolute (ในกรณีนี้คือ dot product) กับ input layer matrix โดยขยับ filter matrix ไปทีละ 2 หน่วย (เรียกว่า stride = 2) จะได้ output ของ layer ที่มีขนาด 3×3×2 เมื่อ 2 คือจำนวนของค่าถ่วงน้ำหนัก

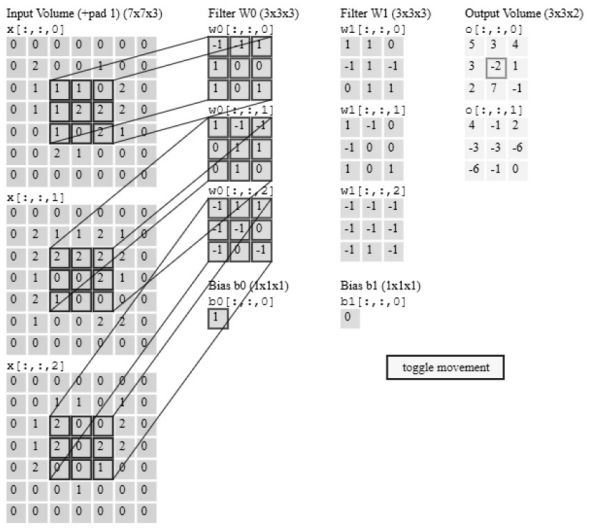

Figure 2 example of convolution layer (Sorce: http://cs231n.github.io/)

1.3 Subsampling layer ชั้นที่ 2 (S2)

รูปที่ 3 แสดงการทำ subsampling แบบ max pool ของ subsampling layer โดยมี filter ขนาด 2x2 หนึ่งตัว และมี stride เท่ากับ 2 subsampling แบบ max pool นี้จะเลือกค่าที่สูงที่สุดออกมา

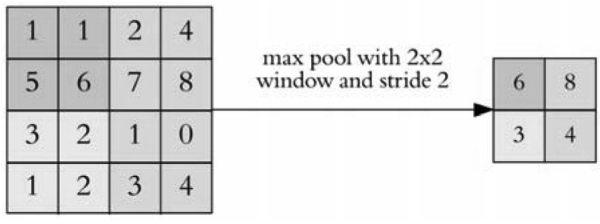

Figure 3 example of subsampling layer (Sorce: http://cs231n.github.io/)

การทำ subsampling เป็นการลด resolution ของ feature map เพื่อลด sensitivity ของเครือข่ายประสาทเทียมแบบคอนโวลูชัน ต่อการขยับและการบิดเบือนของภาพขาเข้ายิ่งขึ้นไปอีก

Subsampling layer ชั้นที่ 2 (S2) ของ LeNet ใน Figure 1 ใช้ filter ทั้งหมด 6 filter แต่ละ filter มีขนาด 2 x 2 มี stride เท่ากับ 2 ดังนั้นจะทำให้เกิด output ทั้งหมด 6 matrix (เรียกว่า feature map) แต่ละ matrix มีขนาด 14 x 14 pixel ใน S2 นี้ 2 x 2 = 4 inputs จะถูกนำมาบวกกัน แล้วคูณด้วย weight ก่อนจะไปนำไปบวกกับค่า bias และเข้าไปสู่ฟังก์ชัน sigmoidal เช่นเดียวกับ Subsampling layer ชั้นที่ 2 (S2) สำหรับ LeNet ใน Figure 1 คือ hidden layer 2 ของเครือข่ายประสาทเทียมทั่วไป โดยมีจำนวน neuron เท่ากับ 6 (เท่ากับจำนวนชุดของ weight) layer นี้มีจำนวน trainable parameter ทั้งหมดเท่ากับ 12 ค่า ((1 weight + 1 bias) x 6 neurons) และมีจำนวน connection ทั้งหมดเท่ากับ 5,880 ครั้ง ((2x2+1) x 14 strides x 14 strides x 6 neurons)

3.1.4 Convolution layer ชั้นที่ 3 (C3)

Convolution layer ชั้นที่ 3 (C3) ของ LeNet ใน Figure 1 ใช้ filter ทั้งหมด 16 filter ด้วยกัน แต่ละ filter มีขนาด 5 x 5 มี stride เท่ากับ 1 จึงทำให้เกิด output ขนาด 10 x 10 pixel ซึ่งข้อแตกต่างของ C3 จาก C1 คือ แต่ละ filter ทำการ convolute กับบาง input feature map เท่านั้นก่อนจะนำผลมารวมกัน ในรูปที่ 4 แต่ละหลัก แสดง input feature map ของ C3 ที่นำมา convolute กับแต่ละ filter การทำเช่นนี้มีเหตุผลสองประการคือ ลดจำนวนการคำนวณของคอมพิวเตอร์ลง และทำให้เกิดการผสมของ feature ทำให้เกิด feature ที่ต่างออกไป

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	X				X	X	X			X	X	X	X		X	X
1	X	X				X	X	X			X	X	X	X		X
2	X	X	X				X	X	X			X		X	X	X
3		X	X	X			X	X	X	X			X		X	X
4			X	X	X			X	X	X	X			X	X	
5				X	X	X			X	X	X	X			X	X

Figure 4 input feature map of C3 (Sorce: LeCun et al. (1998))

เช่นเดียวกับ Convolution layer ชั้นที่ 3 (C3) ของ LeNet ใน Figure 1 คือ hidden layer 3 ของเครือข่ายประสาทเทียมทั่วไป โดยมีจำนวน neuron เท่ากับ 16 (เท่ากับจำนวนชุดของ weight) layer นี้มีจำนวน trainable parameter ทั้งหมดเท่ากับ 1,516 ค่า ((5 x 5 weights x 3 input maps x 6 filters) + (5 x 5 weights x 4 input maps x 9 filters) + (5 x 5 weights x 6 input maps x 1 filter) + 16 biases) และมีจำนวน connection ทั้งหมดเท่ากับ 156,000 ครั้ง (((5x5+1)x3x6 + (5x5+1)x4x9 + (5x5+1)x6x1) x 10 strides x 10 strides)

1.5 Subsampling layer ชั้นที่ 4 (S4)

Subsampling layer ชั้นที่ 4 (S4) ของ LeNet ใน Figure 1 ใช้ filter ทั้งหมด 16 filter แต่ละ filter มีขนาด 2 x 2 มี strideเท่ากับ 2 ดังนั้นจะทำให้เกิด output ทั้งหมด 16 matrix (เรียกว่า feature map) แต่ละ matrix มีขนาด 5 x 5 pixel ซึ่งใน S4 นี้ 2 x 2 = 4 inputs จะถูกนำมาบวกกันแล้วคูณด้วย weight ก่อนจะไปนำไปบวกกับค่า bias

และเข้าไปสู่ฟังก์ชัน sigmoidal เช่นเดียวกับชั้น S2

Subsampling layer ชั้นที่ 4 (S4) สำหรับ LeNet ในรูปที่ 1 คือ hidden layer 4 ของเครือข่ายประสาทเทียมทั่วไปโดยมีจำนวน neuron เท่ากับ 16 (เท่ากับจำนวนชุดของ weight) layer นี้มีจำนวน trainable parameter ทั้งหมดเท่ากับ 32 ค่า ((1 weight + 1 bias) x 16 neurons) และมีจำนวน connection ทั้งหมดเท่ากับ 2,000 ครั้ง ((2x2+1) x 5 strides x 5 strides x 16 neurons)

1.6 Convolution layer ชั้นที่ 5 (C5)

Convolution layer ชั้นที่ 5 (C5) ของ LeNet ในรูปที่ 1 ใช้ filter ทั้งหมด 120 filter ด้วยกัน แต่ละ filter มีขนาด 5 x 5 โดยแต่ละ filter จะ convolute กับ output ของ layer S4 ทั้ง 16 feature maps หรือที่เรียกว่า fully-connected ทำให้เกิด output ทั้งหมด 120 ค่า แต่ละค่ามีขนาด 1 x 1 เช่นเดียวกับ Convolution layer ชั้นที่ 5 (C5) ของ LeNet ใน Figure 1 คือ hidden layer 5 ของเครือข่าย ประสาทเทียมทั่วไป โดยมีจำนวน neuron เท่ากับ 120 (เท่ากับจำนวนชุดของ weight) layer นี้มีจำนวน trainable parameter เท่ากับจำนวน connection เท่ากับ 48,120 ค่า (5 x 5 weights x 16 input maps x 120 filters + 120 biases)

1.7 Fully-connected layer ชั้นที่ 6 (F6)

ชั้นนี้มี 84 neuron ต่อกับ output ทั้ง 120 ค่าจากชั้น C5 แบบ fully connected ชั้นนี้จึงมีจำนวน trainable parameter เท่ากับจำนวน connection เท่ากับ 10,164 ค่า (84 x 120 weights + 84 biases)

1.8 Output layer

ชั้นนี้มี 10 output (แทนตัวเลข 0 ถึง 9) และมี 84 input แต่ละ output คือ Euclidean RBF unit มีค่าเท่ากับ

$$y_i = \sum_j (x_j - w_{ij})^2, \qquad (1)$$

เมื่อ เป็น index ของ output และ เป็น index ของ input โดย ถูกออกแบบในเบื้องต้น ให้เป็นค่า weight ที่แทนตัวเลขที่ถูกต้องของ output การปรับค่า weight ทั้งหมดของเครือข่ายประสาทเทียมแบบคอนโวลูชัน ทำด้วยวิธีการ backpropagation โดย minimize loss function ดังนี้

$$E(W) = \frac{1}{P}\sum_{p=1}^{P} y_{D^p}(Z^p, W), \qquad (2)$$

เมื่อ แทน output ของ class ที่ถูกต้อง

2. Pretrained CNN: GoogLeNet

GoogLeNet ถูกเสนอโดย Szegedy *et al.* (2015) เป็นเครือข่ายประสาทเทียมแบบคอนโวลูชัน (CNN) ที่ถูก pretrain กับรูปภาพใน ImageNet database (www.image-net.org) ในการแข่งขัน ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) รูปภาพที่ใช้ในการ pretrain มีจำนวนหนึ่งล้านรูป แบ่งออกเป็น 1000 ประเภท เช่น คีบอร์ด เมาส์ ดินสอ สัตว์ประเภทต่างๆ

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

Figure 5 layers of GoogLeNet (Sorce: Szegedy *et al*. (2015))

3. Pretrained CNN: Inception-v3

Inception-v3 ถูกเสนอโดย Szegedy et al. (2016) เป็นเครือข่ายประสาทเทียมแบบคอนโวลูชัน (CNN) ที่ถูก pretrain กับรูปภาพใน ImageNet database เช่นเดียวกับ GoogLeNet

Inception-v3 ที่เสนอโดย Szegedy *et al.* (2016) มีชั้นต่างๆ ดังใน Table 1

Type	Patch size/stride or remarks	Input size
Conv	3 x 3/2	299x299x3
Conv	3 x 3/1	149x149x32
Conv padded	3 x 3/1	147x147x32
Pool	3 x 3/2	147x147x64
Conv	3 x 3/1	73x73x64
Conv	3 x 3/2	71x71x80
Conv	3 x 3/1	35x35x192
3 x Inception	As in figure 3	35x35x288
5 x Inception	As in figure 4	17x17x768
2 x Inception	As in figure 6	8x8x1280
Pool	8 x 8	8x8x2048
Linear	Logits	1x1x2048
softmax	Classifier	1x1x1000

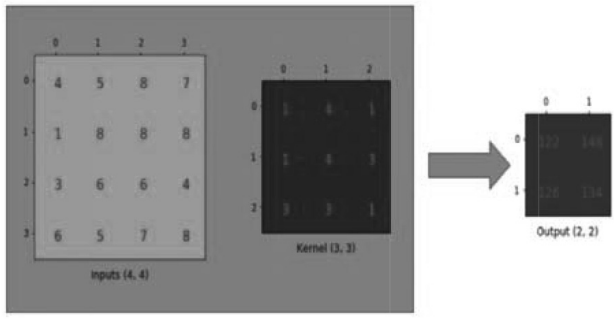

Figure 6 example of convolution (Sorce: Shibuya (2017))

Figure 6 แสดงให้เห็นว่า Convolution layer ทำการ down-sampling ข้อมูลขาเข้า และสามารถถูกนำมาจัดเรียงใหม่ โดยการเติมศูนย์ในบางตำแหน่ง ให้เป็น matrix ขนาด 4x16 เมื่อนำมาคูณกับ input ที่ทำให้เป็นเวกเตอร์ขนาด 16x1 จะได้ output เป็นเวกเตอร์ขนาด 4x1

4. Pretrained CNN: ResNet50

ResNet50 เป็น deep residual learning network โดยมาจากความคิดที่ว่า deeper network (neural network ที่มีหลายชั้น) ยากที่จะเทรนให้ได้ถึงจุด optimum ดังนั้นแทนที่จะพยายามเรียนรู้ฟังก์ชั่นที่ต้องการคือ *H(x)* network นี้จะเรียนรู้ residual function *F(x) = H(x) - x* แทน ดังแสดงใน Figure 7 ResNet50 จึงมีลักษณะเป็น feed-forward network ชนิดหนึ่ง

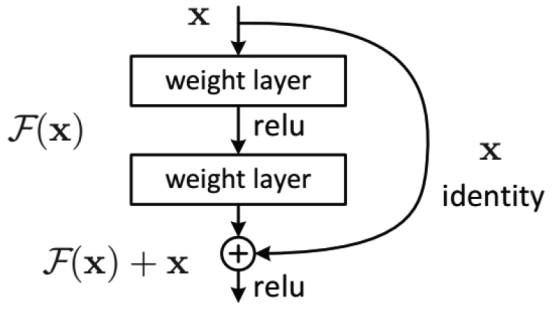

Figure 7 residual function of ResNet50

5. การเรียนรู้แบบถ่ายโอน (transfer learning)

การเรียนรู้แบบถ่ายโอน เป็นการนำ pretrained CNN มาใช้เป็นจุดเริ่มต้นสาหรับ train CNN เพื่อการทำงานในลักษณะอื่น transfer learning รวดเร็วกว่าการเริ่ม train CNN ตั้งแต่ต้นและใช้จำนวนภาพสำหรับ training น้อยกว่ามาก ทั้งนี้เนื่องจาก pretrained CNN ได้เรียนรู้ features ต่าง ๆ ของวัตถุไว้มากแล้ว ตัวอย่างเช่น การเริ่ม train CNN ใหม่ทั้งหมดอาจต้องใช้ภาพเป็นล้าน ๆภาพ ในขณะที่การใช้ transfer learning อาจใช้ภาพเพียงไม่กี่ร้อยภาพ เป็นต้น

Figure 8 แสดงขั้นตอนการทำ transfer learning โดยทั่วไปสาม layer สุดท้ายได้แก่ fully connected layer; softmax layer; และ classification layer จะเป็นชั้นที่เรียนรู้เฉพาะงาน ในขณะที่ชั้นแรก ๆจะเรียน low-level features เช่น ขอบ (edge) หยด (blob) สี (color) การทำ transfer learning จะนำชั้นแรก ๆนี้มาใช้ต่อ ในขณะที่เปลี่ยนสาม layer สุดท้ายให้เป็นเฉพาะสำหรับงานใหม่ การ train จะใช้ภาพเพียงไม่กี่ร้อยภาพ เนื่องจากชั้นแรก ๆได้เรียนรู้ low-level features ไว้แล้ว

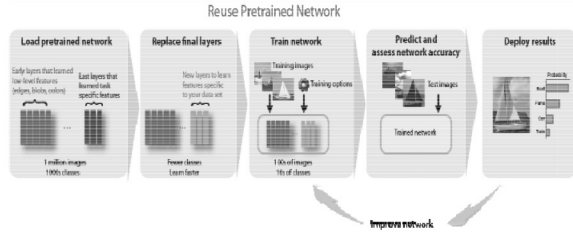

Figure 8 transfer learning (Sorce:: Beale *et al.* (2018a))

วิธีการดำเนินการวิจัยและผลการทดลอง

1. การหา Dataset

งานวิจัยนี้จำแนกอารมณ์ของคนเป็น 7 อารมณ์ด้วยกันคือ ความโกรธ, ความรังเกียจ, ความกลัว, ความสุข, ความเศร้า, ความประหลาดใจและความเป็นกลาง โดยภาพที่ใช้ในการเทรน CNN มาจากสองแหล่งคือจาก Kaggle dataset ที่ใช้ใน Facial Expression Recognition Challenge และ Karolinska Directed Emotional Faces (KDEF) ซึ่งทั้งสองแหล่งมีภาพรวมกันทั้งหมด 40,787 ภาพ โดยนายแบบและนางแบบ มาจากหลากหลายเชื้อชาติ เผ่าพันธุ์ เพศ และ อายุ และภาพเหล่านี้ถูกถ่ายด้วยมุมมองต่าง ๆกัน และใช้โปรแกรม Matlab สำหรับการ convert Kaggle dataset และ Karolinska dataset ให้เป็น image files เพื่อทำการ process ในขั้นต่อไป

โดยลักษณะการแสดงออกทางอารมณ์ของอวัยวะแต่ละส่วนที่นำมาใช้ในการรู้จำ คือ ตา ปาก และใบหน้า

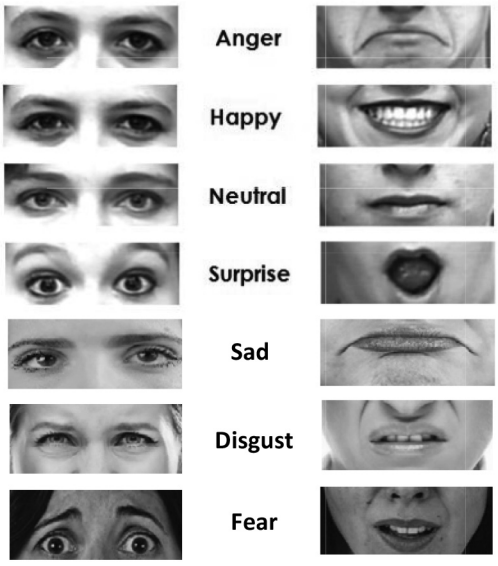

Figure 9 The appearance of the eyes and mouth in an emotional image

2. Pretrained CNN

การ Pretrained CNN สำหรับการจำแนกอารมณ์จากการแสดงออกทางสีหน้าโดยใช้วิธีการ transfer learning ในโปรแกรม matlab จาก pretrained CNN แบบ GoogLeNet, ResNet50 และVGG-19 ดัง Figure 10

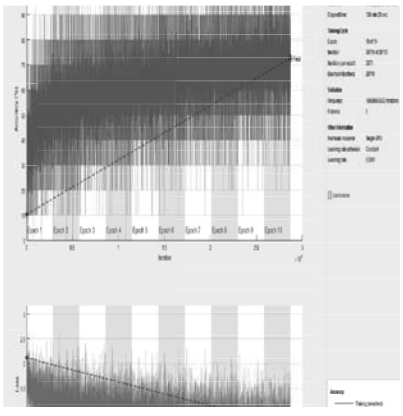

Figure 10 Training progress

3. ผลการทดลอง

ประชากรตัวอย่างจำนวน 30 คน ได้ทำการทดลองใช้โปรแกรมที่คณะผู้วิจัยพัฒนาขึ้น กล้อง webcam หนึ่งตัวได้ถูกตั้งไว้หน้าคอมพิวเตอร์ ได้ผลลัพธ์ดัง Figure 11 โดยแสดงหน้าผู้วิจัยและอารมณ์ที่จับได้ด้วย AI algorithm ที่เสนอด้วยอารมณ์ประหลาดใจ ในรูปพบว่ามีความน่าจะเป็น (probability) ของอารมณ์ประหลาดใจ (surprise) สูงถึง 0.96

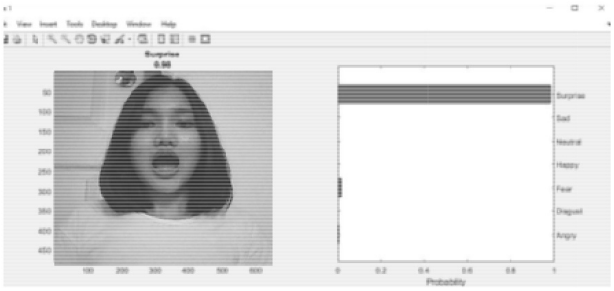

Figure 11 Show the researcher's face and surprise emotion captured by AI algorithm.

ผลการแยกแยะอารมณ์แปลกใจ (surprise) โดยผู้ร่วมทำการทดลองทั้ง 30 คน โดยให้แต่ละคนทำหน้าประหลาดใจหน้ากล้องคนละ 10 ครั้ง และทำหน้าในอารมณ์อื่น 10 ครั้ง เป็นดัง Table 2

Table 2 The results of the surprise experiment

Type of network	Positive accuracy	Probability	Negative accuracy
VGG-19	65%	0.73	75%
ResNet50	80%	0.85	85%
GoogLenet	91%	0.90	97%

Positive accuracy คือความถูกต้องของ AI algorithm ที่ให้ผลเป็นอารมณ์แปลกใจสูงที่สุด เมื่อผู้ทำการทดลองทำหน้าประหลาดใจ

Probability แสดงความน่าจะเป็นเฉลี่ยของอารมณ์ประหลาดใจ เมื่อผู้ทำการทดลองทำหน้าประหลาดใจ

Negative accuracy คือความถูกต้องของ AI algorithm ที่ให้ผลเป็นอารมณ์อื่นๆสูงที่สุดเมื่อผู้ทำการทดลองทำหน้าอารมณ์อื่นๆ

โดยในส่วนของอารมณ์อื่นๆที่ได้ทำการทดสอบพบว่ามีค่าที่ใกล้เคียงกันไม่ห่างกันมาก แต่ในส่วนของความน่าจะเป็นอารมณ์โกรธซึ่งพิจารณาจะเห็นว่า ลักษณะของปากและตา จะคล้ายคลึงกันมากกับอารมณ์ ปกติ จึงทำให้เกิดความสับสันแต่ค่าที่ความน่าจะเป็นที่แสดงออกมีค่าถึง 0.59 ซึ่งถือมีความน่าจะเป็นว่าเกิน 50%

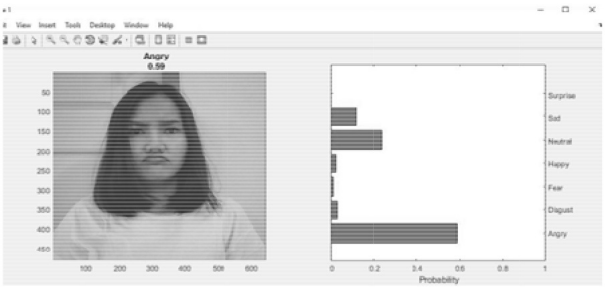

Figure 12 Showing the researcher's face and anger captured by AI algorithm

สรุป

ในงานวิจัยนี้มีวัตถุประสงค์เพื่อออกแบบและพัฒนาระบบตรวจจับอารมณ์ที่แสดงออกทางใบหน้าของมนุษย์ด้วยภาพโดยใช้เทคนิคทางปัญญาประดิษฐ์ เพื่อระบุอารมณ์ความรู้สึกที่สำคัญของมนุษย์ทั้งเจ็ดอารมณ์ ได้แก่ ความโกรธ ความรังเกียจ ความกลัว ความสุข ความเศร้า ความประหลาดใจและความเป็นกลาง ซึ่งได้ผลการทดลองดัง Table 2 จึงสรุปว่าเทคนิคทางปัญญาประดิษฐ์ที่เสนอโดยผู้วิจัยสามารถจำแนกอารมณ์ประหลาดใจของคนได้อย่างถูกต้องแม่นยำ

กิตติกรรมประกาศ

ขอขอบพระคุณท่านรองศาสตราจารย์ ดร. วิทิต ฉัตรรัตนกุลชัย อาจารย์ที่ปรึกษาวิทยานิพนธ์ที่ให้ข้อคิดเห็นและข้อเสนอแนะที่เป็นประโยชน์ยิ่งต่อการวิจัย

เอกสารอ้างอิง

1. Savoiu, A. and J. Wong, Recognizing Facial Expressions Using Deep Learning 2017.
2. Ko, B.C., A Brief Review of Facial Emotion Recognition Based on Visual Information. Sensors (Basel), 2018. 18(2).
3. Tarnowski, P., et al., Emotion recognition using facial expressions 2017.
4. Yanga, D., et al., An Emotion Recognition Model Based on Facial Recognition in Virtual Learning Environment 2017.
5. Fan, Y., et al., Video-based emotion recognition using CNN-RNN and C3D hybrid networks, in Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016. 2016. p. 445-450