บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 อารมณ์และการแสดงออกทางสีหน้า (Emotion and Facial Expression)

อารมณ์ (Emotion) หมายถึงความรู้สึกที่เกิดจากสิ่งเร้าที่เข้ามากระทบ ซึ่งมีทั้งอารมณ์ในทางบวกและทางลบ เช่น ความพึงพอใจและความรู้สึกไม่สมปรารถนา ในแง่ทางประสาทวิทยาอารมณ์มีความหมายคือ การกระทำที่ซับซ้อน (Complex Action Programs) ที่ถูกกระตุ้นจากสิ่งเร้าจากภายนอกหรือสิ่งเร้าภายใน โดยอารมณ์นั้น เป็นส่วนประกอบสำคัญ ของมนุษย์ที่ส่งผลกระทบต่อชีวิตประจำวัน ที่ช่วยกระตุ้นการปฏิสัมพันธ์ทางสังคม ความสนใจ ความเข้าใจ และความทรงจำต่างๆ การตอบสนองทางอารมณ์นั้นประกอบไปด้วย 4 อย่างได้แก่

1. การตอบสนองทางร่างกาย (Bodily Symptoms) เช่น อัตราการเต้นของหัวใจที่เพิ่มขึ้นหรือความร้อนจากผิวหนักที่เพิ่มขึ้น ซึ่งการตอบสนองนี้จะเป็นการตอบสนองแบบอัตโนมัติหรือควบคุมไม่ได้

2. การตอบสนองจากการประพฤติหรือการกระทำ (Action Tendencies) เช่น การตัดสินใจ “สู้หรือหนี” ซึ่งเป็นการตัดสินใจที่เลือกจะหลบเลี่ยงจากเหตุการณ์อันตรายหรือพร้อมสู้กับศัตรูนั้นๆ

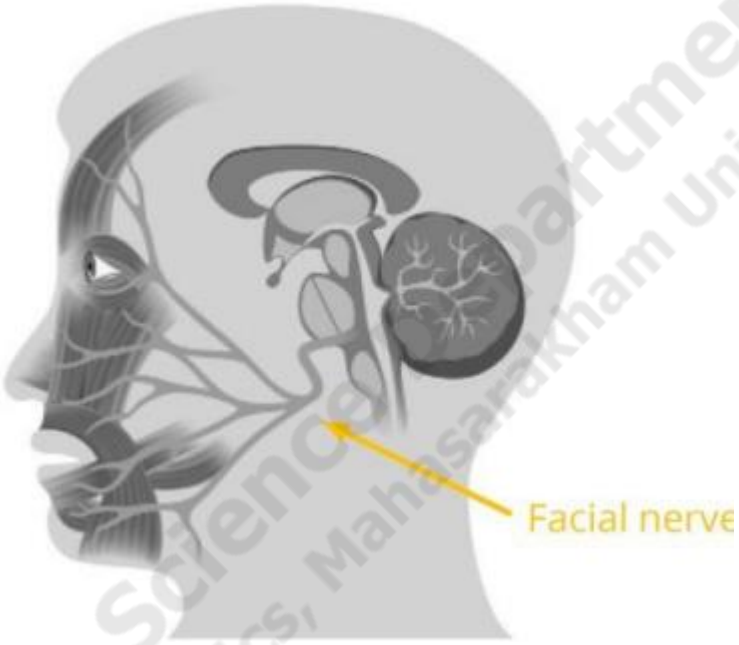
3. การตอบสนองจากการแสดงออกทางสีหน้า (Facial Expression) เช่น การขมวดคิ้วและการเม้มปาก

4. การตอบสนองจากการประเมินการรู้คิด (Cognitive Evaluation) ของเหตุการณ์สิ่งเร้าหรือวัตถุต่างๆ

ในการแสดงออกทางอารมณ์นั้น ส่วนสำคัญที่แสดงอารมณ์ออกมารวดเร็วและมากที่สุดคือส่วนของใบหน้า เช่น เมื่อเรารู้สึกสนุกกับ สิ่งเร้าบางอย่างเราจะยิ้มออกมาทันที หรือหากเราไม่พอใจกับ สิ่งเร้าใดๆ เราก็จะทำหน้าไม่พอใจทันทีซึ่งผู้อื่นจะประมวลผลว่าเรามีอารมณ์ใดจากการสังเกตการเคลื่อนไหวขององค์ประกอบบนใบหน้า ได้แก่ ตา เปลือกตาคิ้วจมูกและปาก

ใบหน้านั้น เป็นส่วนที่มีความซับซ้อนของระบบในการส่งสัญญาณมากที่สุดบนร่างกาย ซึ่งประกอบไปด้วยกล้ามเนื้อต่างๆ มากกว่า 40 มัด และกล้ามเนื้อในแต่ละมัดจะทำงานเป็นอิสระต่อกัน โดยกล้ามเนื้อบนใบหน้านี้ยังเป็นกล้ามเนื้อตำแหน่งเดียวที่กล้ามเนื้อนั้นถูกยึดระหว่างกระดูกกับ เนื้อเยื่อบนใบหน้าหรือถูกยึดแค่บนเนื้อเยื่อเท่านั้นเช่น กล้ามเนื้อบริเวณรอบดวงตาและปากในขณะที่กล้ามเนื้อในส่วนอื่นของร่างกายจะถูกยึดด้วยกระดูกทั้งสองส่วน

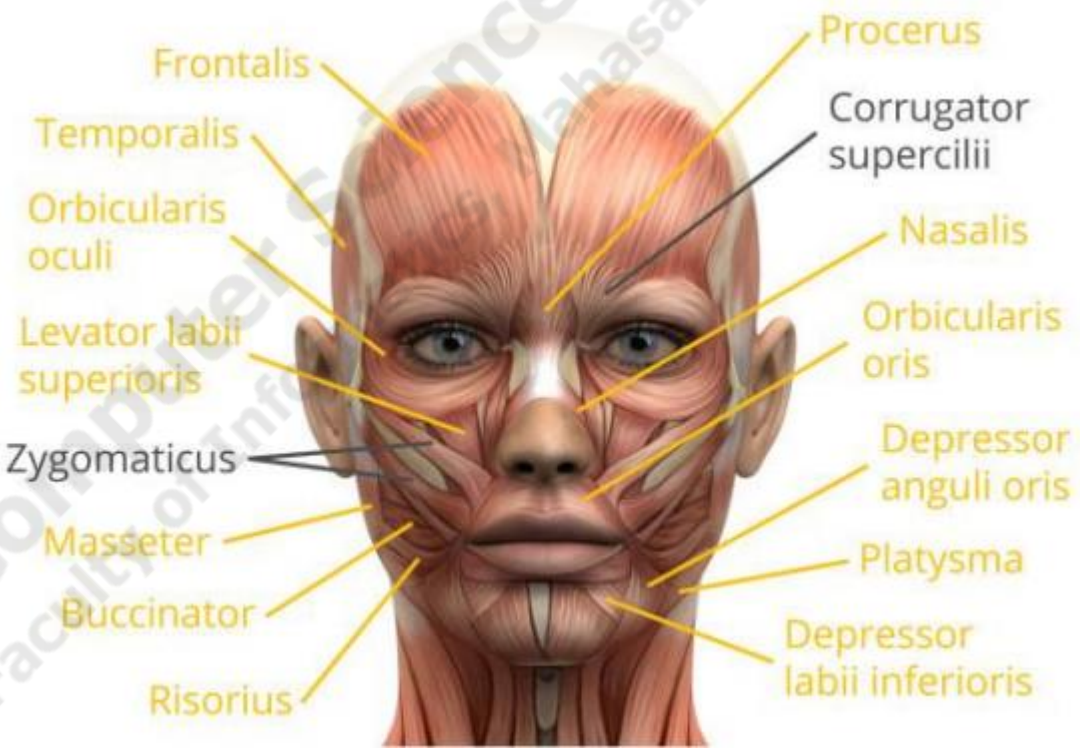
การเคลื่อนที่ของกล้ามเนื้อบนใบหน้านั้นจะถูกกระตุ้นจากเส้นประสาทหนึ่งเส้นคือ Facial Nerve ซึ่งจะเดินทางไปยังไขสันหลังและสมอง โดยระบบประสาทนี้จะมีลักษณะการเคลื่อนที่ 2 ทิศทาง (Bidirectional) คือระบบประสาทสามารถสั่งการให้กล้ามเนื้อเคลื่อนที่จากสัญญาณของสมอง (Brain to Muscle) ในขณะเดียวกันกล้ามเนื้อนั้น จะส่งข้อมูลกลับไปยังสมอง (Muscle to Brain) โดยในทางการแพทย์แล้วเส้นประสาทดังกล่าวจะเรียกว่า “เส้นประสาทสมองคู่ที่ 7 (VII. Cranial Nerve)” ซึ่งเป็นส่วนควบคุมกล้ามเนื้อทั้งหมดที่ใช้ในการแสดงอารมณ์ต่างๆ บนใบหน้า ตัวอย่างดังรูปที่ 2.1



ภาพประกอบที่ 2.1 เส้นประสาทสมองคู่ที่ 7
(ที่มา 55070053.pdf (kmitl.ac.th))

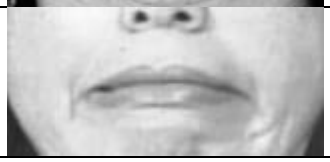
โดยนักจิตวิทยาได้จำแนกอารมณ์ทั่วไป (Universal Facial Expression) ออกเป็น 6 อารมณ์ ได้แก่ รังเกียจ (Disgust) เศร้า (Sadness) สุข (Happiness) กลัว (Fear) โกรธ (Anger) และตกใจ (Surprise) ตัวอย่างแสดงดังรูปที่ 2.2 ซึ่งในแต่ละอารมณ์นั้นจะมีการเคลื่อนไหวของกล้ามเนื้อในตำแหน่งที่แตกต่างกันไป ดังตารางที่ 2.1 และตารางที่ 2.2

ภาพประกอบที่ 2.2 ตัวอย่างการแสดงอารมณ์ทั้ง 6 ประเภท และใบหน้าปกติ (Neutral)
(ที่มา 55070053.pdf (kmitl.ac.th))

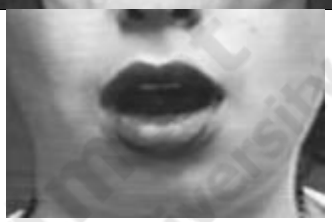


ภาพประกอบที่ 2.3 มัดกล้ามเนื้อต่างๆบนใบหน้า
(ที่มา 55070053.pdf (kmitl.ac.th))

ตารางที่ 2.1 การเคลื่อนไหวและกล้ามเนื้อต่างๆ ที่ใช้ในการแสดงออกทางอารมณ์

หมายเลข	การเคลื่อนไหว	กล้ามเนื้อที่ทำงาน	ภาพตัวอย่าง
1	คิ้วด้านในยกสูงขึ้น	Frontalis, Pars Medialis	
2	คิ้วด้านนอกยกสูงขึ้น	Frontalis, Pars Lateralis	
3	คิ้วยกต่ำลง	Corrugator Supercilii, Depressor Supercilii	
4	เปลือกตาบนยกสูงขึ้น	Levator Palpebrae Superioris	
5	แก้มยกสูงขึ้น	Orbicularis Oculi, Pars Orbitalis	
6	เปลือกตาหรี่แคบลง	Orbicularis Oculi, Pars Palpebralis	
7	ย่นจมูก	Levator Labii Superioris Alaquae Nasi	
8	มุมปากยกสูงขึ้น	Zygomaticus Major	
9	มุมปากลดต่ำลง	Depressor Anguli Oris	
10	ริมฝีปากล่างยกต่ำลง	Depressor Labii Inferioris	
11	ริมฝีปากยืดเป็นแนวตรง	Risorius, Platysma	

ตารางที่ 2.1 การเคลื่อนไหวและกล้ามเนื้อต่างๆ ที่ใช้ในการแสดงออกทางอารมณ์(ต่อ)

หมายเลข	การเคลื่อนไหว	กล้ามเนื้อที่ทำงาน	ภาพตัวอย่าง
12	ริมฝีปากชิดกันแน่น	Orbicularis Oris	
13	ขากรรไกรตก	Masseter, Relaxed Temporalis and Internal Pterygoid	

ตารางที่ 2.2 การทำงานของกล้าเนื้อในอารมณ์ต่างๆ

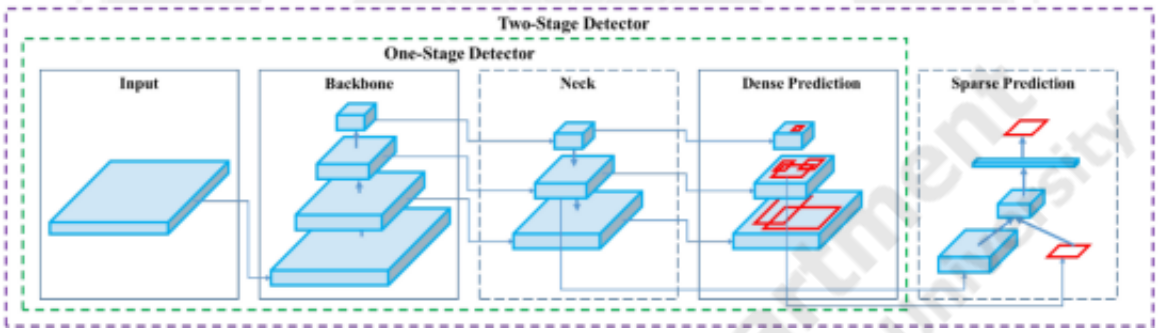
อารมณ์	หมายเลขของกล้ามเนื้อที่ทำงาน
สุข (Happiness)	5 + 8
เศร้า (Sadness)	1 + 3 + 9
ตกใจ (Surprise)	1 + 2 + 4 + 13
กลัว (Fear)	1 + 2 + 3 + 4 + 6 + 11 +13
โกรธ (Anger)	3 + 4 + 6 + 12
รังเกียจ (Disgust)	7 + 9 + 10

2.1.2 ระบบตรวจจับวัตถุ

การตรวจจับวัตถุ (Object Detection) [5] เป็นการตรวจจับและระบุวัตถุภายในภาพนิ่งหรือวิดีโอซึ่งทำการวิเคราะห์ผ่านคอมพิวเตอร์ ในยุคปัจจุบันด้วยความทันสมัยของเทคโนโลยีที่มากขึ้นทำให้มีการหันมาเลือกใช้เทคนิคที่ใช้ Deep Learning เพื่อให้รวดเร็วและแม่นยำมากยิ่งขึ้น ซึ่งกระบวนการของการตรวจจับจะเริ่มจากการนำเข้าข้อมูล และทำการดึงเอาคุณลักษณะที่สำคัญของชุดข้อมูลฝึกฝนมาประมวลผลเพื่อนำไปแยกประเภทของวัตถุนั้นแล้วก็ทำนายผลลัพธ์สุดท้ายออกมาว่าจัดอยู่ใน Class ชนิดใด โดยเทคนิคที่ใช้สำหรับตรวจจับวัตถุจะสามารถแบ่งออกเป็น 2 ประเภทหลัก ๆ ดังนี้

1) One-Stage Object Detection จะมีจุดเด่นในเรื่องของความเร็วที่ใช้ในการประมวลผล ซึ่งปัจจุบันมีการพัฒนารุ่นใหม่ออกมาขึ้นเรื่อย ๆ ซึ่งเพิ่มความแม่นยำที่สูงมากและสามารถนำไปใช้งานจริงได้ เช่น YOLO, YOLOv2, YOLOv3, YOLOv4 และ SSD (Single Shot multi-Box Detector) ซึ่งในส่วนของอัลกอริทึมตระกูล YOLO ก็ได้มีการพัฒนารุ่น tiny ซึ่งมีความรวดเร็วที่เพิ่มขึ้นกว่าตัวแบบตั้งต้นในแต่ละเวอร์ชั่น แต่จะทำให้ความแม่นยำลดลงตามไปด้วย

2) Two-Stage Object Detection เป็นกระบวนการแรก ๆ ที่ได้มีการใช้และเป็นที่นิยมก่อนที่ One-Stage จะถูกคิดค้นขึ้นมา โดยจะใช้ Region Proposal ซึ่งจะมีจุดเด่นในเรื่องของความแม่นยำที่สูงแต่จะไม่รวดเร็วเนื่องจากมีกระบวนการทำงานหลายขั้นตอน เช่น R-CNN, Fast R-CNN และ Faster R-CNN



ภาพประกอบที่ 2.4 สถาปัตยกรรมการตรวจจับวัตถุ

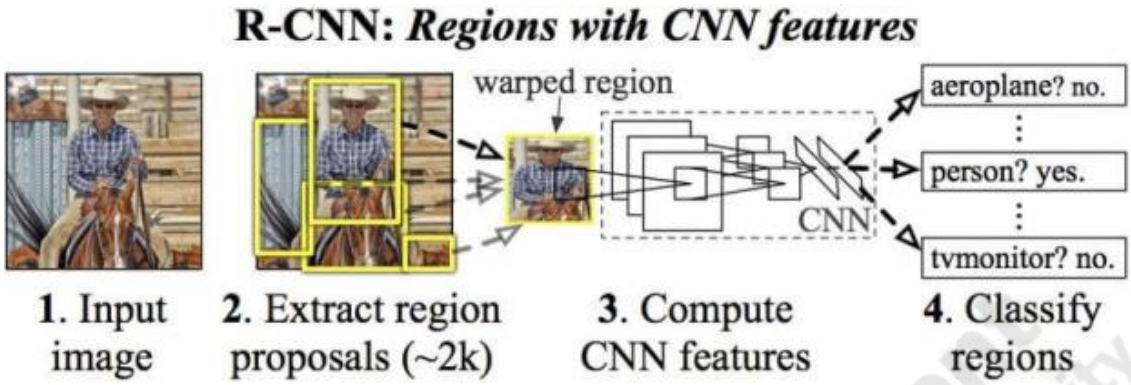
(ที่มา Two concepts of architectural object detection [6]. The YOLOv5 network... | Download Scientific Diagram (researchgate.net))

2.1.2.1 R-CNN

โดย R-CNN [5] เป็นหนึ่งในตัวแรกของ Deep Learning ที่ใช้ในการตรวจจับวัตถุและเป็นตัวอย่างของเครื่องจักรที่ตรวจจับด้วย 2 ขั้นตอน ดังนี้

1) ใช้ R-CNN เป็นตัวตรวจจับวัตถุที่ต้องใช้อัลกอริทึม เช่น Selective Search หรือเทียบเท่าเพื่อเป็นตัวเลือกในการสร้างขอบเขตที่อาจมีวัตถุอยู่ในจุดนั้น ๆ (Girshick, Donahue,Darrell, & Malik, 2014) โดยการนำ Region Proposal จำนวนประมาณ 2,000 แปลงเป็นกรอบขอบเขต

2) จากนั้นนำเข้าสู่โครงข่ายประสาทเทียม โดย CNN ทำการจำแนกประเภทคุณลักษณะและเลเยอร์ต่างๆ โดยผลลัพธ์ที่ออกมาจะประกอบไปด้วยคุณลักษณะที่ดึงออกมาได้จากภาพหรือวิดีโอและ feature ที่แยกออกมาแล้วนั้นจะถูกนำไปเข้า SVM เพื่อทำการจำแนกประเภท โดยปัญหาในวิธีการมาตรฐานของ R-CNN นั้นคือประมวลผลช้าไม่ได้เป็นเครื่องมือในการตรวจจับวัตถุแบบครบวงจรที่สมบูรณ์

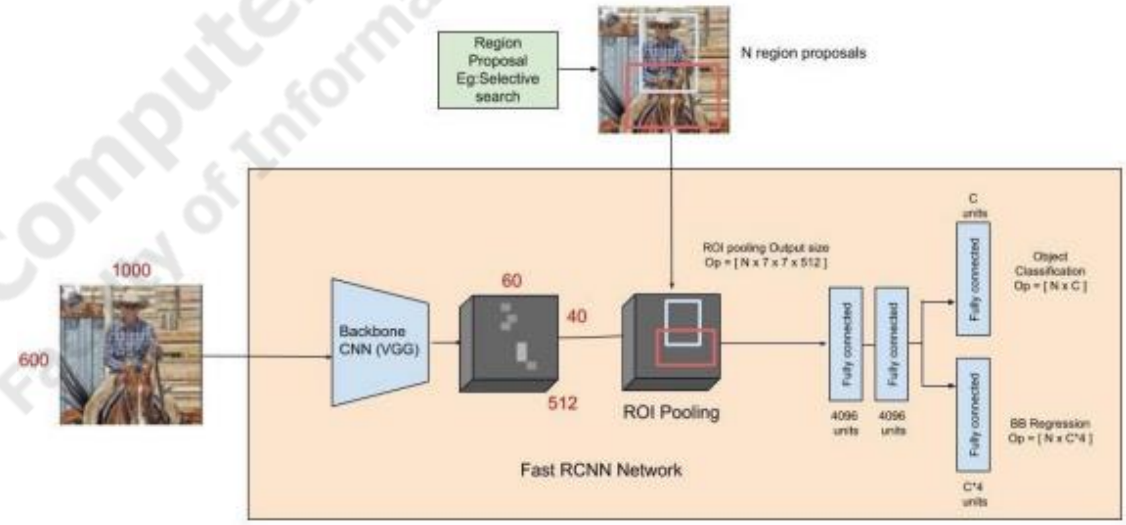


ภาพประกอบที่ 2.5 กระบวนการทำงาน R-CNN

(ที่มา Introduction to object detection and Evolution: RCNN, Fast-RCNN, Faster-RCNN, YOLO | by Sumeet Sewate | Medium)

2.1.2.2 Fast R-CNN กับ Faster R-CNN

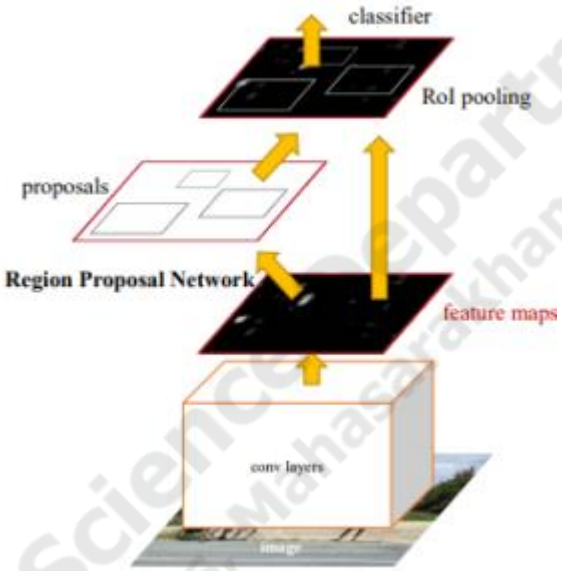
จากนั้นจึงได้มีการพัฒนา Fast R-CNN (Girshick, 2015) ทำให้ R-CNN กลายเป็นเครื่องมือในการตรวจจับวัตถุที่ใช้ Deep Learning แบบ End-To-End ด้วยการกำจัดข้อกำหนดการค้นหาโดยการแทนที่ข้อจำกัดส่วนนั้นด้วยการใช้ Region Proposal Network (RPN) ซึ่งมีหน้าที่ในการสกัดคุณลักษณะที่มีความน่าจะเป็นว่าเป็นวัตถุที่มาจาก Feature Map และทำการเปลี่ยนจากการนำคุณลักษณะที่แยกออกมาได้จากรูปภาพนำเข้า SVM เป็นการต่อเติมโครงข่ายประสาทเทียมแทนที่โดยยังมีปัญหาในเรื่องของความเร็วซึ่งสามารถรองรับได้เพียง 5 FPS บน GPU



ภาพประกอบที่ 2.6 กระบวนการทำงาน Fast R-CNN

(ที่มา Introduction to object detection and Evolution: RCNN, Fast-RCNN, Faster-RCNN, YOLO | by Sumeet Sewate | Medium)

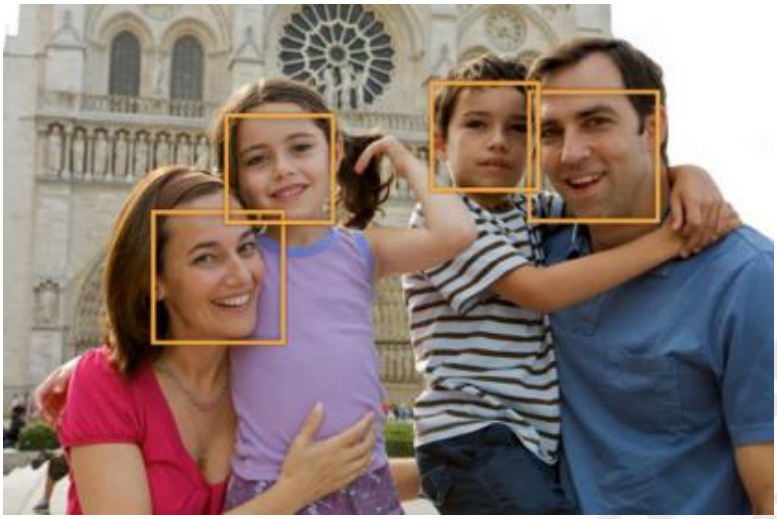
หลังจากนั้นไม่นานก็มีการพัฒนา Faster R-CNN โดยแนวคิดหลักคือการนำเอาโครงข่ายประสาทเทียมมารวมเข้ากับ Selective Search ซึ่งได้มีการสร้าง Anchor Box จำนวนหนึ่งเนื่องจากมีการนำโครงข่ายประสาทเทียมมากำหนดหรือทาย จำนวน Region และ RPN จะทำการทายตำแหน่งและขนาดของ Anchor Box รวมถึงความน่าจะเป็นที่จะเป็นกรอบล้อมวัตถุ ซึ่งกระบวนการทั้งหมดจะแยกการทำออกมาจากกันและทำไปทีละขั้นตอน โดยการพัฒนาในครั้งนี้ทำให้มีความเร็วเพิ่มขึ้นถึง 10 เท่าเมื่อทำการเทียบกับ R-CNN (Ren, He, Girshick, & Sun, 2017)



ภาพประกอบที่ 2.7 กระบวนการทำงาน Faster R-CNN
(ที่มา R-CNN vs Fast R-CNN vs Faster R-CNN - A Comparative Guide (analyticsindiamag.com))

2.1.3 การตรวจจับใบหน้า (Face Detection)

การตรวจจับใบหน้า [5] คือกระบวนการในการหาพื้นที่ของใบหน้าบนรูปภาพ ในปัจจุบันมีแอพพลิเคชั่นมากมายที่ต้องการใช้รูปใบหน้าเพื่อนำไปพัฒนาระบบ เช่น ระบบรู้จำใบหน้า (Face-Recognition) ระบบรู้จำอารมณ์บนใบหน้า (Facial Expression Recognition) ระบบรู้จำองค์ประกอบบนใบหน้า (Facial Attribute Recognition) และระบบการประกอบโครงสร้างใบหน้าขึ้นมาใหม่(Facial Shape Reconstruction) โดยทุกระบบจะต้องใช้การตรวจจับใบหน้าเป็นขั้นตอนแรกในการประมวลผล ทำให้การตรวจจบนใบหน้านั้นจะต้องมีประสิทธิภาพที่ดีที่สุด กล่าวคือความผิดพลาดในการตรวจจับ สิ่งอื่นที่ไม่ใช่ใบหน้าต้องน้อย (False Positive) ความถูกต้องในการตรวจจับใบหน้าต้องสูง (True-Positive) และการประมวลผลต้องไว เพื่อให้การประมวลผลในขั้น ตอนต่อๆ ไปให้ผลลัพธ์ที่ดีที่สุด



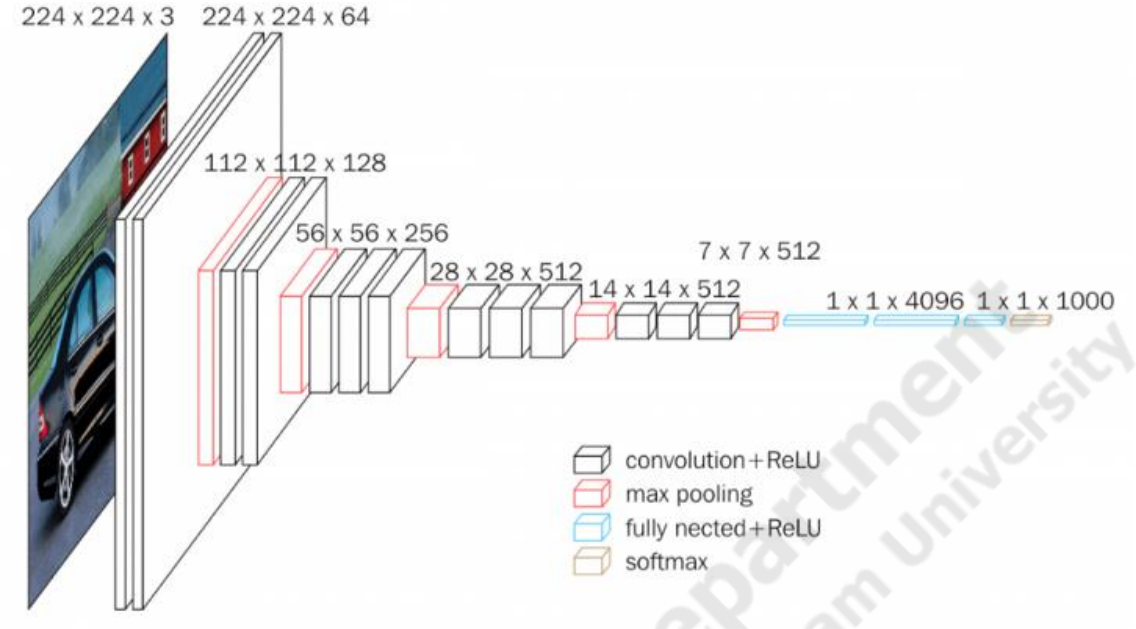
ภาพประกอบที่ 2.8 ตัวอย่างการตรวจจับใบหน้า
(ที่มา Apple concept for biometric facial recognition could hint at 'iPhone 8' | AppleInsider)

2.1.4 การรู้จำอารมณ์บนใบหน้า (Facial Expression Recognition)

ระบบรู้จำอารมณ์บนใบหน้านั้นถูกพัฒนาขึ้นเพื่อใช้ทดสอบผลกระทบจากความพึงพอใจจากสินค้าและบริการ และจากวัตถุทางกายภาพ เช่น อาหาร บรรจุภัณฑ์อาหาร ภาพเคลื่อนไหวในวีดิโอ ภาพนิ่ง เสียงกลิ่น และสิ่งเร้าที่สัมผัสได้ที่มีความเกี่ยวข้องทางอารมณ์และการตอบสนองบนใบหน้า โดยเป้าหมายหลักของการวิเคราะห์อารมณ์นั้นคือการหาการเปลี่ยนแปลงทางอารมณ์ที่ถูกแสดงออกมาโดยอัตโนมัติเช่น การที่เปลือกตาขยายกว้างขึ้น เป็นกุญแจสำคัญที่สะท้อนให้เห็นถึงการเปลี่ยนแปลงทางอารมณ์ที่ถูกกระตุ้นโดนสิ่งเร้าภายนอกหรือมโนภาพที่เกิดขึ้นโดยในปัจจุบันระบบรู้จำอารมณ์นั้นได้ถูกนำไปใช้ในหลายสาขา

2.1.5 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) [5] หรือ โครงข่ายประสาทแบบคอนโวลูชันเป็นจำลองการมองเห็นของมนุษย์ที่มองเป็นส่วนย่อยๆ และนำกลุ่มของส่วนย่อย ๆ มาผสานกัน เพื่อดูว่าสิ่งที่เห็นอยู่คืออะไร โดยใช้ค่าพิเซลที่ได้จากข้อมูลอินพุต มีทั้งหมด 3 สี ได้แก่ สีแดง, น้ำเงิน, และเขียว สามารถใช้เลข 0 ถึง 255 เพื่อแทนค่าความเข้มของสี



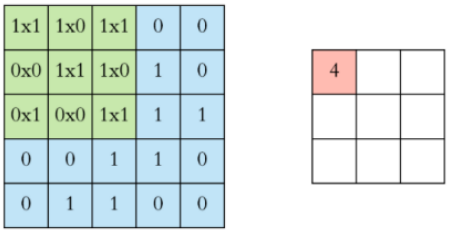
ภาพประกอบที่ 2.9 Convolutional Neural Network VGG-16
(ที่มา VGGNet-16 Architecture: A Complete Guide | Kaggle)

รับภาพ Input เข้ามาเป็น array ขนาด 224*224*n โดยที่ n คือจำนวนโหนด Depth จากนั้นทำ Max Pooling เพื่อหาค่าที่มากที่สุดของจุดภาพด้วยตัวกรอง(filter) ขนาด 3*3 คือการลดขนาดของจุดภาพโดยที่ให้สูญเสียรายละเอียดของภาพน้อยที่สุด แล้วก็ประมวลผลแบบนี้ไปเรื่อย ๆ ตามจำนวนของโหนด Hidden layer ไปจนถึงชั้น Fully Connected (FC) ส่งค่าคำนวณจากโหนดหนึ่ง ไปอีกโหนดหนึ่ง เชื่อมต่อ Hidden Layer ต่างๆ เข้าด้วยกัน แล้วทำการประมวลผล Output ออกมาตาม Class จากนั้น เมื่อถึงชั้น Soft Max ก็จะแปลงค่าของ output ออกมาให้อยู่ในรูปแบบความน่าจะเป็น

โครงสร้างของ Convolutional Neural Network ประกอบได้ดังนี้

2.1.5.1 Convolutional

เป็น Layer หลักของ CNN ทำหน้าที่รับ Input เข้ามา แปลงภาพให้เป็นพิกเซล ที่กำหนดให้เป็น 0 ถึง 255 จากนั้นจะใช้การดำเนินการทางคณิตศาสตร์เพื่อหาคุณสมบัติที่สำคัญจากรูปภาพเตอร์ การคำนวณจะเริ่มจากการกำหนดค่าในตัวกรอง (filter) หรือ เคอร์เนล (kernel) ที่ช่วยดึงคุณลักษณะที่ใช้ในการรู้จำวัตถุออกมา หรือที่เรียกว่า Feature Map



ภาพประกอบที่ 2.10 Feature Map (Natthawat Phongchit, 2561)

การทำงานของ CNN จะทำการ Sliding Windows (Filter) เพื่อค้นหาองค์ประกอบของภาพ เช่น สี หรือรูปร่าง ทำได้ด้วยสมการ 2.1

$$output\ of\ size = \frac{N-F+2P}{S} + 1 \qquad (2.1)$$

โดยที่ N คือ ขนาดของภาพ

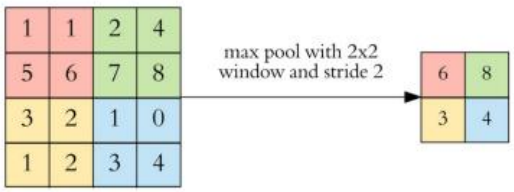
F คือ ขนาดของ Filter

P คือ จำนวนของ Padding

S คือ จำนวนของ Stride (จำนวนของการขยับ Filter)

2.1.5.2 Pooling Layer

Pooling Layer เป็นชั้นที่เชื่อมจาก Convolutional Layer โดยมีเป้าหมายคือทำให้ขนาดของ Feature Map ลดลงด้วยการหาค่าเฉลี่ย (Average Pooling) หรือหาค่าที่สูงที่สุด (Max Pooling) และจะเลื่อนตัวกรองไปตาม Stride ที่กำหนดไว้ โดยขนาดตัวกรองของการทำ Max Pooling นิยมเรียกกันว่า Pool Size



ภาพประกอบที่ 2.11 Pooling Layer (Natthawat Phongchit, 2561)

2.1.5.3 Fully Connected layer

โดยขั้นตอนการหาค่าแต่ละโหนด ในขั้นตอน Fully Connected layer สามารถทำได้ด้วยสมการ 2.2

$$H_i = \sum_{i=o}^{n-1}(x_i \cdot W_i) \quad (2.2)$$

โดยที่ H_i คือ ผลลัพธ์ Hidden Layer โหนดที่ i

n คือ จำนวน Input ของโหนดก่อนหน้า

x_i คือ ข้อมูลของโหนด Input

W_i คือ ค่าน้ำหนัก

และเมื่อได้ผลลัพธ์นำข้อมูลเข้าฟังก์ชันที่รับผลรวมการประมวลผลทั้งหมด Sigmoid Function

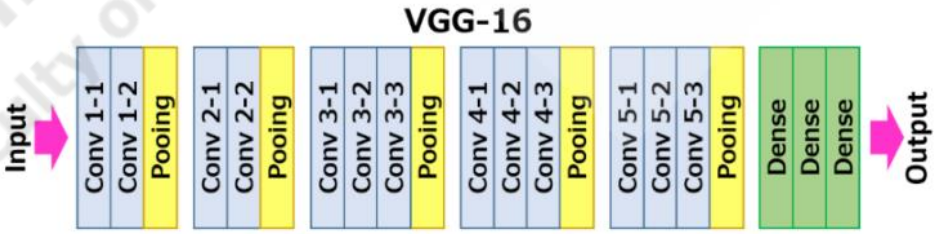
$$F(A) = \frac{1}{1+e^{-A}} \quad (2.3)$$

โดยที่ F คือ ผลลัพธ์ Sigmoid Function มีค่าระหว่าง 0 ถึง 1

A คือ ผลลัพธ์ของ Hidden Layer

ในงานวิจัยฉบับนี้ใช้ VGG16

VGG16 ในรูปแบบโครงข่ายประสาทเทียมที่เสนอโดย K. Simonyan และ A. Zisserman จาก University of Oxford แบบจำลองนี้มีความแม่นยำในการทดสอบสูงสุด 5 อันดับแรก 92.7% ใน ImageNet ซึ่งเป็นชุดข้อมูล ของรูปภาพมากกว่า 14 ล้านภาพที่เป็นของที่มีชื่อเสียงที่ส่งถึง ILSVRC-2014 มันทำให้การปรับปรุงให้ทำงานดีกว่า AlexNet โดยแทนที่ตัวกรองเคอร์เนลขนาดใหญ่ (11 และ 5 ในชั้นแรกและครั้งที่สองตามลำดับ) ด้วยตัวกรองขนาดเคอร์เนล 3 × 3 ดังภาพที่ 2.12



ภาพประกอบที่ 2.12 โครงสร้าง VGG-16

ตารางที่ 2.3 Summary of VGG16 Architecture

Layer		Feature Map	Size	Kernel Size	Stride	Activation
Input	Image	1	224 x 244 x3	-	-	-
1	2 X Convolution	64	224 x 224 x64	3x3	1	relu
	Max Pooling	64	112 x 112 x 64	3x3	2	relu
3	2 X Convolution	128	112 x 112 x 128	3x3	1	relu
	Max Pooling	128	56 x 56 x 128	3x3	2	relu
5	2 X Convolution	256	56 x 56 x 256	3x3	1	relu
	Max Pooling	256	28 x 28 x 256	3x3	2	relu
7	3 X Convolution	512	28 x 28 x 512	3x3	1	relu
	Max Pooling	512	14 x 14 x 512	3x3	2	relu
10	3 X Convolution	512	14 x 14 x 512	3x3	1	relu
	Max Pooling	512	7 x 7 x 512	3x3	2	relu
13	FC	-	25088	-	-	relu
14	FC	-	4096	-	-	relu
15	FC	-	4096	-	-	Relu
Output	FC	-	1000	-	-	Softmax

เลเยอร์ชั้นที่หนึ่งและสอง: อินพุตเป็นภาพ RGB 224x224x3 ซึ่งผ่านเลเยอร์ Convolutional ชั้นที่หนึ่งและที่สองพร้อมกับ 64 Feature Map หรือ Filter ที่มีขนาด 3 x 3 และการรวมขนาดเดียวกันโดยมี Stride เท่ากับ 14 ขนาดของรูปภาพเปลี่ยนเป็น 224x224x64 จากนั้น VGG16 จะใช้เลเยอร์ร่วมสูงสุดหรือเลเยอร์การสุ่มตัวอย่างย่อยด้วยขนาด ตัวกรอง 3 × 3 และก้าวหนึ่งของทั้งสองขนาดภาพที่ได้จะลดลงเป็น 112x112x64

เลเยอร์ชั้นที่สามและสี: ถัดไปมีสองเลเยอร์ Convolutional พร้อมกับฟีเจอร์ 128 แผนที่ที่มีขนาด 3 × 3 และทำ Stride ที่ 1 จากนั้นจะมีเลเยอร์รวมสูงสุดอีกครั้งที่มีขนาดตัวกรอง 3 × 3 และทำ stride ที่ 2 เลเยอร์นี้เหมือนกับเลเยอร์รวมก่อนหน้านี้ยกเว้นมี Feature Map คุณสมบัติ 128 ชุดดังนั้นผลลัพธ์จะลดลงเหลือ 56x56x128

เลเยอร์ชั้นที่ห้าและหก: เลเยอร์ที่ห้าและหกเป็นเลเยอร์ Convolutional ที่มีขนาดตัวกรอง 3 x 3 และก้าวหนึ่งทั้งสองใช้แผนที่ฟีเจอร์ 256 รายการเลเยอร์ Convolutional สองเลเยอร์จะตามด้วยเลเยอร์สูงสุดรวมกับขนาดตัวกรอง 3 × 3, ทำ stride ของ 2 และมี 256 feature map

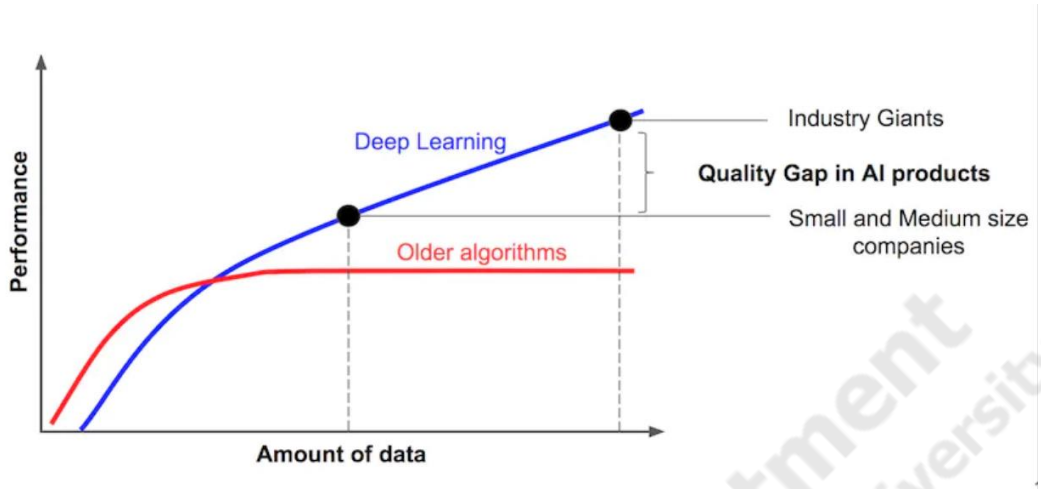
เลเยอร์ชั้นที่เจ็ดถึงชั้นสิบสอง: ถัดไปคือชุด Convolutional 3 เลเยอร์สองชุดตามด้วยชั้นการรวมสูงสุดเลเยอร์ Convolutional ทั้งหมดมีตัวกรอง 512 ขนาด 3 × 3 และก้าวหนึ่งขนาดสุดท้ายจะลดลงเป็น 7x7x512

เลเยอร์ชั้นที่สิบสาม: เอาท์พุทเลเยอร์ Convolutional จะแผ่ผ่านเลเยอร์ที่เชื่อมต่ออย่างเต็มที่ด้วย Feature Map 25088 แต่ละขนาด 1 × 1

เลเยอร์ที่สิบสี่และสิบห้า: ถัดไปเป็นอีกสองเลเยอร์ที่เชื่อมต่ออย่างสมบูรณ์ด้วย 4096 หน่วยเลเยอร์เอาท์พุท: ในที่สุดก็มี ชั้นเอาท์พุท SoftMax กับ 1,000 ค่าที่เป็นไปได้

2.1.6 การขยายข้อมูล (Data Augmentation)

การขยายข้อมูลเป็นการเพิ่มจำนวนข้อมูลให้มีจำนวนมากขึ้นเพื่อให้เพียงพอต่อการนำไปใช้ในการทำงานกับรูปภาพการขยายข้อมูลหมายถึงการเพิ่มจำนวนของรูปภาพในฐานข้อมูล ในการทำงานกับข้อมูลในลักษณะปกติจะหมายถึงการเพิ่มแถวข้อมูล การขยายข้อมูลถูกนำเนื่องจากมนุษย์มีจำนวนข้อมูลที่จำกัด แต่ตามหลักการแล้วยิ่งมีข้อมูลมากขึ้นโมเดลของ Machine Learning ก็จะมีประสิทธ์ภาพดีขึ้น อย่างไรก็ตามการประมวลผลข้อมูลในทุกแบบย่อมเกี่ยวพันกับค่าใช้จ่าย ค่าใช้จ่ายนี้อาจจะเป็นในรูปแบบเงินตรา การลงแรงของมนุษย์ หรือพลังการประมวลผลของคอมพิวเตอร์ และแน่นอนว่าจะต้องเสียเวลาในการประมวลผล ดังนั้นเราจึงต้องการการขยายข้อมูลที่มีอยู่แล้วเพื่อเพิ่มจำนวนของข้อมูลที่จะป้อนให้โมเดล Machine Learning เพื่อให้โมเดลสามารถทำหน้าที่ได้อย่างมีประสิทธิภาพ



ภาพประกอบที่ 2.13 ความสัมพันธ์ระหว่างปริมาณของข้อมูลในการฝึกสอนและความแม่นยำของตัวแบบการเรียนรู้เชิงลึก

(ที่มา Big challenge in Deep Learning: training data | HackerNoon)

การขยายข้อมูลสามารถทำได้หลายวิธีในการทำงานกับรูปภาพจะใช้ การหมุนรูปภาพเดิม เปลี่ยนสภาพแสงในภาพ ครอบตัดภาพให้ลักษณะต่างออกไป ดังนั้นภาพหนึ่งภาพสามารถสร้างเป็นข้อมูลภาพที่แตกต่างกันหลายๆ ภาพได้ตามเทคนิคที่ใช้ในการขยายข้อมูล ด้วยวิธีนี้เองเราสามารถลดปัญหาการ Overfit ของโมเดล Machine Learning กล่าวคือปัญหาที่ตัวแบบทำงานได้แม่นยำมากกับข้อมูลรูปภาพที่ใช้ในการฝึก แต่ทำงานได้ไม่แม่นยำในข้อมูลจริงซึ่งเป็นข้อมูลที่ตัวแบบไม่เคยเรียนรู้มาก่อน ในทางกลับกันถ้าใช้เทคนิคที่มีการ over sampling เช่น SMOTE จะทำให้มีโอกาสที่จะเกิด Overfit ซึ่งเป็นสิ่งที่ควรหลีกเลี่ยง

2.2 งานวิจัยที่เกี่ยวข้อง

งานวิจัยของ ธนพล พุ่มลำเจียก, 2016 [1] เรื่อง " FACIAL EXPRESSION RECOGNITION FROM VIDEO SEQUENCE USING LOCAL GABOR FILTERS AND PCA PLUS LDA " [1] เป็นการวิจัยที่ทำการรู้จำอารมณ์บนใบหน้าจากวีดีโอ โดยใช้ตัวกรองกาบอร์ วิธีการวิเคราะห์ องค์ประกอบหลักและการวิเคราะห์จำแนกประเภทเชิงเส้น โดยผลการทดลองเมื่อทาการเปรียบเทียบค่าความแม่นยำ ที่ได้จะพบว่าในฐานข้อมูล CK+ อัลกอริทึมที่ถูกพัฒนาขึ้นมีความแม่นยำ 94.58% ซึ่งมากกว่าการใช้การวิเคราะห์องค์ประกอบหลัก และการวิเคราะห์จำแนกประเภทเชิงเส้น อยู่ 10.16% และ 11.08% ตามลำดับ และใน ฐานข้อมูล JAFFE อัลกอริทึมที่ถูกพัฒนาขึ้นมีความแม่นยำ 97.5% ซึ่งมากกว่าการวิเคราะห์ องค์ประกอบหลักและการวิเคราะห์จำแนกประเภทเชิงเส้น อยู่ 7.63% และ 14.37% ตามลำดับ และค่าความแม่นยำในการจำแนกระหว่างอารมณ์ปกติและอารมณ์โกรธ อัลกอริทึมที่พัฒนาขึ้นมีความแม่นยำ 95% ซึ่งมากกว่า 20%

งานวิจัยของ จุฑามาศ มาบรรดิษ และ คณะ, 2016 [2] เรื่อง “การรู้จำอารมณ์บนใบหน้า โดยใช้วิธีการวิเคราะห์องค์ประกอบหลัก และการวิเคราะห์จำแนกประเภทเชิงเส้น” เป็นงานวิจัยพัฒนาระบบรู้จำอารมณ์บนใบหน้าเพื่อต้องการให้ระบบดังกล่าวมีประสิทธิภาพและมีความถูกต้องมากยิ่งขึ้น อีกทั้งยังต้องการให้ระบบดังกล่าวมีความแพร่หลายมากขึ้นในประเทศไทย เนื่องจากกระบวนการในการวิเคราะห์อารมณ์นั้น ยังมีความซับซ้อนในการพัฒนา ผลการทดลองได้ทำการทดลองการรู้จำอารมณ์บนใบหน้าโดยใช้ PCA และ LDA ทดสอบกับฐานข้อมูลสามฐานข้อมูล คือ JAFFE, CK และ CPEKPS ปรากฏว่า PCA และ LDA ให้ค่าความถูกต้องที่ ใกล้เคียงกันมาก โดย PCA ให้ผลที่ดีกว่า LDA เล็กน้อย และ L2-norm ให้ผลดีกว่า L1-norm

งานวิจัยของ Keyur Patel และ คณะ, 2020 [3] เรื่อง “Facial Sentiment Analysis using AI” เป็นงานวิจัยนำเสนอการสำรวจอย่างเป็นระบบโดยละเอียดเพื่อวิเคราะห์วิธีการที่ทันสมัยในปัจจุบันสำหรับการจดจำอารมณ์ใบหน้าในภาพนิ่งและพารามิเตอร์ต่าง ๆ ที่มีอิทธิพลต่อผลลัพธ์ของวิธีการเหล่านี้ เราได้พัฒนาระบบภาษีตามวิธีการที่แตกต่างกันที่ใช้สำหรับการตรวจจับใบหน้าการสกัดคุณลักษณะและการจำแนกอารมณ์ ผลการทดลองเราได้เปรียบเทียบวิธีการตรวจจับการสกัดและการจำแนกประเภทต่างๆและสรุปว่าวิธีการใดมีความโดดเด่นมากขึ้นในการบรรลุประสิทธิภาพที่ดีขึ้นในพลังการคำนวณที่มีอยู่ โดยการหารือเกี่ยวกับปัญหาและการวิจัยในปัจจุบันความท้าทายในอนาคตเราสรุปว่ายังมีการวิจัยที่จำเป็น

ตารางที่ 2.4 ตารางการเปรียบเทียบระบบงานที่เกี่ยวข้อง

ระบบงาน / ฟังก์ชั่นการทำงาน	FACIAL EXPRESSION RECOGNITION FROM VIDEO SEQUENCE	Facial Expression Recognition Based on PCA and LDA	A Sentimental Analysis on Facial Expression Recognition	Facial emotion analysis from video
ตรวจจับใบหน้า	/	/	/	/

ตารางที่ 2.4 ตารางการเปรียบเทียบระบบงานที่เกี่ยวข้อง(ต่อ)

ระบบงาน / ฟังก์ชั่นการทำงาน	FACIAL EXPRESSION RECOGNITION FROM VIDEO SEQUENCE	Facial Expression Recognition Based on PCA and LDA	A Sentimental Analysis on Facial Expression Recognition	Facial emotion analysis from video
คัดแยกตำแหน่งขององค์ประกอบบนใบหน้า	/		/	
จดจำบนใบหน้า	/	/	/	/
จำแนกอารมณ์ต่างๆ	/	/	/	/
ความแม่นยำของแต่ละอารมณ์	/		/	/