บทความวิชาการ

บทความปริทัศน์: งานวิจัยการรู้จำการกระทำของมนุษย์ด้วยการเรียนรู้เชิงลึก

ธนกร สว่างโลก* และ ปกป้อง ส่องเมือง
สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ ศูนย์รังสิต

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 0 2564 4441 อีเมล: tanakon.sawa@dome.tu.ac.th
DOI: 10.14416/j.kmutnb.2022.07.013
รับเมื่อ 8 มีนาคม 2564 แก้ไขเมื่อ 7 มิถุนายน 2564 ตอบรับเมื่อ 30 กรกฎาคม 2564 เผยแพร่ออนไลน์ 27 กรกฎาคม 2565

บทคัดย่อ

การที่คอมพิวเตอร์สามารถรู้จำการกระทำของมนุษย์เกิดจากการที่โมเดลสามารถทำนายได้ว่า มนุษย์กำลังกระทำการกระทำอะไรอยู่ โดยใช้ข้อมูลวิดีโอของสถานะการกระทำในปัจจุบัน การรู้จำการกระทำของมนุษย์สามารถนำไปประยุกต์ได้หลายด้าน เช่น ความปลอดภัย ความบันเทิง การอำนวยความสะดวก ซึ่งงานวิจัยในสาขานี้มีการเปลี่ยนแปลงรวดเร็วมาก ดังนั้น ในงานสำรวจนี้รวบรวมงานวิจัยการรู้จำของมนุษย์ที่ใช้เทคนิคการเรียนรู้เชิงลึกซึ่งมีประสิทธิภาพสูง ถูกนำไปใช้ต่อยอดในงานวิจัยอื่น หรือนำไปใช้ในอุตสาหกรรม โดยจะนำเสนอความท้าทายในงานรู้จำการกระทำ โจทย์ปัญหาของผู้วิจัย เทคนิคและสถาปัตยกรรมที่ใช้ ข้อจำกัดของงานวิจัย

คำสำคัญ: การรู้จำการกระทำ การเรียนรู้เชิงลึก ความท้าทายในงานรู้จำการกระทำ

การอ้างอิงบทความ: ธนกร สว่างโลก และ ปกป้อง ส่องเมือง, "บทความปริทัศน์: งานวิจัยการรู้จำการกระทำของมนุษย์ด้วยการเรียนรู้เชิงลึก," *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 33, ฉบับที่ 3, หน้า 1–12, เลขที่บทความ 233-274840, ก.ค.–ก.ย. 2566.

Academic Article

Human Action Recognition in Deep Learning Aspect: Review Article

Tanakon Sawanglok* and Pokpong Songmuang
Department of Computer Science, Faculty of Science and Technology, Thammasat University, Rangsit Centre, Pathum Thani, Thailand

* Corresponding Author, Tel. 0 2564 4441, E–mail: tanakon.sawa@dome.tu.ac.th
DOI: 10.14416/j.kmutnb.2022.07.013
Received 8 March 2021; Revised 7 June 2021; Accepted 30 July 2021; Published online: 27 July 2022

Abstract

Computer capability of human action recognition is on the basis that the model can make and inference-based inductive predictions on human actions using video action reasoning of the current activity status. Human action recognition can be widely applied in several fields such as security, entertainment, and facilitation. The research in this field is changing rapidly. In this review article, research information on human action recognition using deep learning techniques will be gathered. State-of-the-art techniques will be further applied in various fields and across industries. The review article presents challenges in action recognition, researcher problems, techniques, research methodology algorithm along with research limitations.

Keywords: Action Recognition, Deep Learning, Challenges in Action Recognition

Please cite this article as: T. Sawanglok and P. Songmuang, "Human action recognition in deep learning aspect: Review article," *The Journal of KMUTNB*, vol. 33, no. 3, pp. 1–12, ID. 233-274840, Jul.–Sep. 2023 (in Thai).

1. บทนำ

การรู้จำการกระทำของมนุษย์ (Human Action Recognition) เป็นหัวข้อที่ได้รับความสนใจจากนักวิจัยทั่วโลก เพราะสามารถไปประยุกต์ใช้กับหลากหลายสายงานหรือแอปพลิเคชัน (Application) ได้แก่

1) การตรวจระวังเกี่ยวกับการเห็น (Visual Surveillance) [1] เช่น ตรวจเช็คว่ามีบุคคลแปลกหน้า หรือไม่ได้รับอนุญาตเข้ามาในเขตที่อยู่อาศัย แล้วสามารถส่งสัญญาณเตือนไปที่เจ้าหน้าที่รักษาความปลอดภัย การมีระบบที่ตรวจจับการกระทำจากกล้องวงจรปิดช่วยลดความเสี่ยงที่เกิดอันตรายกับบุคคล และเพิ่มความปลอดภัยขึ้น

2) การค้นคืนวิดีโอ (Video Retrieval) [2] เช่น การค้นหาวิดีโอที่มีเนื้อหา หรือการกระทำที่คล้ายกัน เพื่อแนะนำให้ผู้ใช้ได้มีประสิทธิภาพมากขึ้น

3) อุตสาหกรรมบันเทิง (Entertainment) [3] เช่น เกมที่ใช้เทคนิครู้จำการกระทำหรือการเคลื่อนไหวมาใช้ในเกม เช่น เกมเต้น ตีปิงปอง ต่อสู้ โยนโบว์ลิ่ง ซึ่งการใช้ร่างกายและเล่นได้ทั้งครอบครัวพร้อมกัน

4) รถยนต์ขับขี่อัตโนมัติ (Autonomous Driving Vehicle) [4] ซึ่งมีระบบที่ขับเคลื่อนแบบอัตโนมัติไม่ต้องใช้มนุษย์ และใช้การรู้จำการกระทำในการตรวจจับรถยนต์อื่น คนเดินเท้า ซึ่งช่วยลดอุบัติเหตุในท้องถนน

การกระทำของมนุษย์ (Human Action) ครอบคลุมตั้งแต่การเคลื่อนไหวอย่างง่าย เช่น การเคลื่อนไหวของมือ เท้า เช่น โบกมือ หยิบสิ่งของ ไปจนถึงการเคลื่อนไหวซับซ้อน ใช้หลายอวัยวะพร้อมกัน หรือมีวัตถุมาเกี่ยวข้อง เช่น ว่ายน้ำ การตีกอล์ฟ ซึ่งการกระทำเหล่านั้นมีช่วงของการกระทำไม่เท่ากัน บางการกระทำอาจใช้เวลาแค่ 4 วินาที บางการกระทำอาจใช้เวลา 20 นาที

บทความปริทัศน์นี้รวบรวมงานวิจัยในอดีตเกี่ยวกับการรู้จักการกระทำ (Action Recognition) วิเคราะห์ และนำเสนอในรูปแบบ เบส อิวิเดนซ์ รีวิว (Best Evidence Review) ซึ่งศึกษางานวิจัยที่สนใจโดยดูจากวิธีการในแต่ละกลุ่มและผลลัพธ์ของแต่ละงาน ผู้วิจัยนำเสนอโดยจัดกลุ่มของงานวิจัยในอดีตตามลักษณะที่มีร่วมกัน และยกงานวิจัยที่น่าสนใจ ได้ประสิทธิภาพสูงมาอธิบายและวิเคราะห์

งานวิจัยการตรวจจับการกระทำในอดีตแบ่งออกเป็น 2 แนวทางหลัก ได้แก่

1) การรู้จำการกระทำด้วยวิธีสกัดคุณลักษณะแบบดั้งเดิม (Traditional Approach) ที่ใช้มนุษย์ในการค้นหาแบบแผน มีข้อจำกัดคือ บางครั้งจำกัดใช้ได้เฉพาะในโดเมนเท่านั้น และใช้เวลาในการค้นหาแบบแผนนาน

2) การรู้จำการกระทำด้วยการเรียนรู้เชิงลึก (Deep Learning Approach) ที่ใช้เทคนิคการเรียนรู้ของเครื่องจักร (Machine Learning) มาใช้ในการหาแบบแผน หรือคุณลักษณะเด่นของข้อมูลได้แบบอัตโนมัติ ข้อจำกัด คือ ถ้าโมเดล หรือข้อมูลมีความซับซ้อนมาก ทำให้การเรียนรู้ให้โมเดลมีประสิทธิภาพสูงได้ยาก และล่าช้า

ผู้เขียนสนใจไปในแนวทางที่ 2 ที่ใช้การเรียนรู้เชิงลึกใช้ในการทำการรู้จักการกระทำ ซึ่งแนวทางนี้แบ่งออกเป็น 3 กลุ่มย่อย ได้แก่

1) มัลติเพิล สตรีม (Multiple Streams) โดยสมาชิกในกลุ่มมองเห็นปัญหาว่าคอนโวลูชัน 2D (Convolution2D) ไม่สามารถสกัดคุณลักษณะแบบสปาเชียลเทมโพรอล (Spatiotemporal) ซึ่งเป็นข้อมูลที่มีลักษณะเชิงพื้นที่ของรูปภาพและความเปลี่ยนแปลงในแต่ละเฟรมอยู่ด้วยกัน จากปัญหาข้างต้นสมาชิกในกลุ่มจึงมีแนวคิดแยกฝึกฝน (Training) ข้อมูลลักษณะเชิงพื้นที่ (Spatial) ซึ่งใช้ข้อมูลภาพสี RGB กับข้อมูลการเคลื่อนไหว (Temporal) ซึ่งใช้ข้อมูลวิดีโอภาพการเคลื่อนไหว (Optical Flow) เพื่อเพิ่มประสิทธิภาพ

2) สเปซไทม์ (Space-time) โดยสมาชิกในกลุ่มต้องการใช้วิธีการที่เป็นเจเนริก (Generic Approach) และสามารถสกัดคุณลักษณะแบบสปาเชียลเทมโพรอลได้ จึงได้ประยุกต์ใช้คอนโวลูชัน 3D กับสถาปัตยกรรมหลายชนิด

3) วิธีที่ใช้เฉพาะเจาะจง (Specific Propose) โดยสมาชิกในกลุ่มต้องการแก้ปัญหาหลายแบบ เพื่อเพิ่มประสิทธิภาพให้มากขึ้นในปัญหาที่ตัวเองสนใจ ซึ่งมีรายละเอียดในแต่ละหัวข้อของงานสำรวจงานวิจัยการรู้จักการกระทำ

1.1 ความท้าทายในการทำการรู้จักการกระทำ

งานวิจัยการรู้จักการกระทำมีจุดที่มักจะผิดพลาด หรือ

ได้ค่าความผิดพลาดสูง (Error) ที่มาจากปัจจัยเหล่านี้

1.1.1 ความแตกต่างภายในกลุ่มและระหว่างกลุ่ม (Intra- and Inter-Class Variations)

การกระทำของมนุษย์มีความท้าทายหลายอย่างซึ่งปัจจัยเหล่านั้น คือ

1) ท่าทางการเคลื่อนไหว การกระทำอย่างเดียวกัน คน 2 คน อาจใช้ท่าท่าทางที่ไม่เหมือนกัน เช่น การว่ายน้ำ มีหลายท่าทางการว่าย ท่าผีเสื้อ ท่ากบ ท่าฟรี

2) มุมกล้องที่ใช้บันทึกการกระทำ มุมที่ใช้จับท่าทางมีหลายมุมกล้อง เช่น ถ่ายแบบบุคคลที่ 1 หรือบุคคลที่ 3 ถ่ายมุมกล้องกดลงมา มุมเสยขึ้น ถ่ายด้านหน้าคน ถ่ายด้านข้างคน

3) ความคล้ายกันของการกระทำ การกระทำบางอย่างมีลักษณะที่คล้ายกัน เช่น การเดินและการวิ่ง การเดินขึ้นบันไดและการเดินลงบันได

1.1.2 พื้นหลังที่ไม่จำเป็นและการเคลื่อนไหวของกล้อง (Cluttered Background and Camera Motion)

งานวิจัยการรู้จำการกระทำ บางเทคนิคสามารถให้ผลดีในชุดข้อมูลควบคุม (Control Dataset) แต่ได้ผลที่ไม่น่าพอใจสำหรับชุดข้อมูลที่ไม่ผ่านการควบคุม (Uncontrol Dataset) ความผิดพลาดนี้เป็นผลมาจากปัจจัยเหล่านี้

1) ภาพพื้นหลังของผู้ทำการกระทำ พื้นหลังของชุดข้อมูลที่ไม่ได้ถูกควบคุมมักประกอบไปด้วยคลื่นรบกวน (Noise) หรือสิ่งที่ไม่เกี่ยวข้องกับการกระทำ เช่น การวิ่งในที่สาธารณะ อาจมีคนที่ทำการกระทำอื่นปะปนไปด้วย ซึ่งการสกัดคุณลักษณะจากทั้งรูปภาพเหล่านี้อาจมีผลทำให้ประสิทธิภาพโดยรวมลดลงได้

2) การกำหนดการตั้งค่าของกล้อง วิดีโอที่บันทึกการกระทำ แต่ละวิดีโออาจมีการตั้งค่าความเข้มแสง (Illumination Conditions) หรือมุมมองที่ต่างกัน (Viewpoint Changes) เช่น ภาพที่มาจากกล้อง Closed-circuit Television (CCTV) ซึ่งบันทึกไว้ทั้งวันอาจมีค่าแสงที่แตกต่างกันในแต่ละวัน

1.1.3 ข้อมูลที่มีการระบุประเภทมีไม่เพียงพอ (Insufficient Annotated Data)

การเก็บข้อมูลภาพหรือวิดีโอ การกระทำของมนุษย์และเป้าหมายที่ต้องการรู้จำ (Class Label, Target) เป็นงานที่ใช้ทรัพยากรมนุษย์ รวมถึงเวลาสูงมาก นอกจากนี้การเก็บข้อมูล Dataset ด้วยมนุษย์อาจมีการให้ Label ผิดได้ ซึ่งทำให้ประสิทธิภาพ Model โดยรวมลดลงด้วย จากปัญหาข้างต้น นอกจากการเพิ่มจำนวน Dataset และ Label ให้ถูกต้องมากขึ้น งานวิจัยการรู้จำการกระทำของมนุษย์ควรเรียนรู้จากข้อมูลที่มี Label และข้อมูลที่ไม่มี Label ได้อย่างมีประสิทธิภาพ เช่น มีการใช้เทคนิค Semi-supervised Learning [5]

1.1.4 การให้ความสำคัญของความสามารถในการทำนาย

ขั้นตอนวิธี (Algorithm) ที่ใช้ในการรู้จำการกระทำของมนุษย์มักเรียนรู้โดยให้ความสำคัญทุกเฟรมภาพเท่ากัน ซึ่งในความเป็นจริงแล้ว เฟรมภาพที่ใช้ในการคัดแยกประเภท (Classify) อาจจะมีแค่บางเฟรมที่สำคัญ ทำให้เฟรมที่เหลือมีความสำคัญน้อยและมีจำนวนมากทำให้ใช้เวลาในการเรียนรู้มากขึ้น และทำให้ประสิทธิภาพลดลงได้

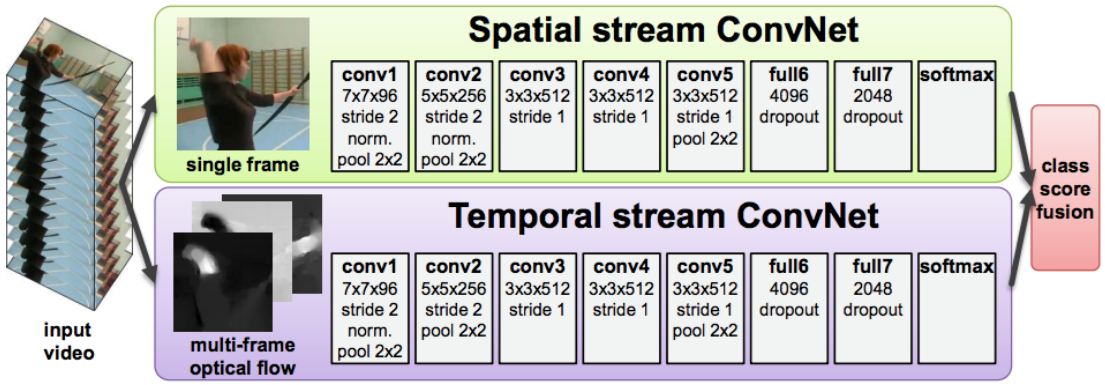
จากการสำรวจงานวิจัยการรู้จำการกระทำของมนุษย์ด้วยการเรียนรู้เชิงลึก ร่วมด้วยการวิเคราะห์ข้อดี ข้อจำกัดของแต่ละกลุ่มแล้ว พบว่า นอกจากจะใช้ในงานรู้จำการกระทำของมนุษย์แล้ว สามารถนำไปประยุกต์ใช้กับงานที่มีลักษณะเป็นเวลาต่อเนื่อง (Time-series) หรือลำดับ (Sequence) ได้อีกด้วย เช่น งานประมวลผลธรรมชาติ งานประมวลผลสัญญาณ หรือการพยากรณ์อากาศ

งานสำรวจงานวิจัยการรู้จำการกระทำของมนุษย์ด้วยการเรียนรู้เชิงลึก มีโครงสร้างงานดังนี้ ส่วนที่ 1 บทนำ อธิบายภาพรวมของงานสำรวจ ส่วนที่ 2 งานสำรวจงานวิจัยรู้จำการกระทำ อธิบายงานวิจัยในแต่ละกลุ่ม และส่วนที่ 3 Conclusion สรุปงานสำรวจ

2. งานสำรวจงานวิจัยการรู้จำการกระทำ

การรู้จักการกระทำมี 2 ส่วนประกอบ ได้แก่

1) คุณลักษณะสำหรับรู้จำการกระทำ (Action Feature) ซึ่งได้จากการสกัดคุณลักษณะ (Extract Feature) จากข้อมูลดิบ (Raw Data) เช่น ภาพ (Image) วิดีโอ (Video)



รูปที่ 1 สถาปัตยกรรมแบบ Two Stream [6]

เสียง (Sound) ระเบียน (Record) โดยใช้เทคนิคการสกัดคุณลักษณะ เช่น คอนโวลูชัน เลเยอร์ (Convolutional Layer) ซึ่งเป็นกระบวนการสกัดคุณลักษณะของข้อมูลออกมาแบบอัตโนมัติ โดยทำการสุ่มเคอร์เนล (Kernel) ขนาดต่างกัน นำไปทำการคูณเมทริกซ์ (Matrix Multiplication) กับข้อมูลที่มีลักษณะเป็นอาเรย์ที่มีหลายมิติ (Multi-dimensional Array) นอกจากนี้ยังมีพูลลิง เลเยอร์ (Pooling Layer) ที่สามารถลดขนาดข้อมูลลงได้ ผลลัพธ์ที่ได้คือ คุณลักษณะที่เป็นตัวแทนจากข้อมูลดิบอยู่ในรูปของเวกเตอร์ (Vector) หรือเมทริกซ์ ซึ่งนำไปใช้กับส่วนที่

2) การรู้จำการกระทำทำหน้าที่สร้างโมเดลที่ทำนาย Label หรือ Class จาก Feature โดยใช้เทคนิคการเรียนรู้ของเครื่องจักร เช่น โครงข่ายประสาทเทียม (Neural Network) ซัพพอร์ตเวกเตอร์แมทชีน (Support Vector Machine) หรือข่ายงานเบย์ (Bayesian Network)

ในส่วนต่อจากนี้อภิปรายเกี่ยวกับองค์ประกอบของการรู้จักการกระทำในแง่มุมของการเรียนรู้เชิงลึก โดยนำเสนอเฉพาะงานวิจัยที่น่าสนใจใน พ.ศ. 2558–2562 จากงานประชุมวิชาการ

2.1 สถาปัตยกรรมที่ใช้ในการรู้จำการกระทำ (Architecture in Action Recognition)

ผลลัพธ์จากงานแข่งขันรู้จำรูปภาพขนาดใหญ่ที่มีชื่อเสียง หรืออิมเมจเน็ต (The ImageNet Large Scale Visual Recognition Challenge) ทำให้นักวิจัยทั่วโลกสนใจในการเรียนรู้เชิงลึกมากขึ้น ในวงการการรู้จักการกระทำมีการนำการเรียนรู้เชิงลึกมาประยุกต์ใช้ ในอดีตก่อน พ.ศ. 2557 นักวิจัยทำการเรียนรู้เชิงลึกจัดการงานการรู้จักการกระทำโดยใช้เครือข่ายสตรีมเดียว (Single Stream Network) ซึ่งเป็นการใช้ 1 เครือข่ายที่มีคอนโวลูชันหลาย Layer และใช้เทคนิคในการจัดการนำเข้าหลายแบบ เช่น

1) ซิงเกิลเฟรม (Single Frame) คือ การที่ประมวลผลทีละ 1 เฟรม

2) เลทฟิวชัน (Late Fusion) คือ การประมวลผลที่ 2 จุดเฟรม เช่น ประมวลผลเฟรมที่หัวและท้ายเฟรมของวิดีโอโดยใช้ 2 Architecture ที่ Share Parameters ร่วมกัน

3) เออรีฟิวชัน (Early Fusion) คือ การประมวลผลที่ช่วงของเฟรม

4) สโลว์ฟิวชัน (Slow Fusion) คือ การประมวลผลที่ช่วงของเฟรม และมีการรวมข้อมูลที่หลายสเตจของเครือข่าย

ใน พ.ศ. 2557 Simonyan และ Zisserman [6] ได้นำเสนอ Two-Stream Convolutional Networks สำหรับงานการรู้จักการกระทำผู้วิจัยมีแนวคิดว่าการแยก Training ข้อมูลลักษณะเชิงพื้นที่กับข้อมูลการเคลื่อนไหวให้ผลลัพธ์ดีมากขึ้น วิธีการคือใช้ข้อมูลนำเข้า 2 แบบ ดังรูปที่ 1 แบบที่ 1 คือ ข้อมูลวิดีโอภาพสี RGB (RGB-video) สำหรับ Spatial Stream แบบที่ 2 คือ วิดีโอภาพการเคลื่อนไหวสำหรับ Temporal Stream จากนั้น Training ข้อมูลแยกกัน 2 Stream

ผลลัพธ์คือความน่าจะเป็นของแต่ละ Class และเลือกผลลัพธ์สุดท้ายจากคะแนนเฉลี่ย 2 Stream

งานวิจัยชิ้นนี้นำเสนอ Architecture ใหม่ในการทำการรู้จักการกระทำ ซึ่งต่อมาเป็นฐานสำหรับนำไปต่อยอดในงานวิจัยอื่น แต่มีข้อจำกัดที่ต้องคำนวนภาพวิดีโอการเคลื่อนไหวก่อนสำหรับทุก Dataset

จากงานวิจัยและเทคนิคข้างต้นเป็นฐานสำหรับพัฒนาในหลายงานวิจัยด้านการรู้จักการกระทำ ใน พ.ศ. 2558–2562

งานวิจัยในการรู้จักการกระทำที่ใช้การเรียนรู้เชิงลึกแบ่งออกเป็น 3 กลุ่มย่อย ได้แก่ 1) Multiple Streams 2) Space-time และ 3) Specific Propose โดยเริ่มจากกลุ่ม Multiple Streams

2.2 มัลติเพิล สตรีม (Multiple Streams)

งานวิจัยในกลุ่มที่ 1 Multiples Stream เป็นกลุ่มที่ต่อยอดแนวคิดมาจาก Two-streams Architectures โดย Wang และคณะ [7] นำเสนอ Architecture ใหม่ ชื่อว่า Temporal Segment Network ใน พ.ศ. 2559 โดยงานวิจัยนี้เกิดขึ้นเพื่อแก้ไขปัญหาความเชื่อมโยงในระยะยาว (Long Term Dependency) เนื่องจากในงานวิจัยในอดีตตอน Training ส่วนมากแบ่งช่วงของวิดีโอเป็นช่วงสั้น (Short Snippet) โดย Training และ Prediction ไปทีละช่วง ซึ่งการกระทำบางประเภทต้องใช้เวลาในการที่จะบอกได้ว่าเป็นการกระทำใด เช่น การวิ่งเตะลูกฟุตบอล การแบ่งข้อมูลเป็น Short Snippet ทำให้ Model ขาดข้อมูล Long Term Dependency ไป

Feichtenhofer และคณะ [8] นำเสนองานวิจัยใน พ.ศ. 2559 ที่หาวิธีพัฒนาประสิทธิภาพของ Two Stream Architecture โดยค้นพบว่า การรวมคุณลักษณะของข้อมูลเชิงพื้นที่ และข้อมูลการเคลื่อนไหวเข้าด้วยกัน (Fusion Technique) ทำให้ประสิทธิภาพสูงขึ้นได้ มีวิธีการ Fusion ที่ศึกษาทั้งหมด 5 แบบ ได้แก่ การฟิวชันแบบรวม (Sum Fusion) การฟิวชันแบบค่าสูงสุด (Max Fusion) การฟิวชันแบบต่อเข้าด้วยกัน (Concatenate Fusion) การฟิวชันแบบไบลิเนียร์ (Bilinear Fusion) และการฟิวชันแบบคอนโวลูชัน (Conv Fusion) รวมถึงศึกษาการ Fusion ที่ Layer ต่างกัน จากการ Fusion Spatial และ Temporal เข้าด้วยกัน ทำให้ Architecture ใหม่สามารถหา Spatiotemporal Features ได้ และผลการศึกษาพบว่า การ Fusion ที่ Last Convolution เช่น Convolution Layer ที่ 5 รวมถึงวิธี Fusion แบบ Sum และ Convolution ให้ประสิทธิภาพที่ดีที่สุด

Girdhar และคณะ [9] พบปัญหาว่า การกระทำบางอย่างเป็นการกระทำพื้นฐานสำหรับการกระทำอื่น เช่น การยิงบาสเก็ตบอล มีการกระทำพื้นฐานคือ การเดินหรือวิ่ง ซึ่งพบเจอได้ในการเล่นฟุตบอล หรือกีฬาชนิดอื่น ซึ่งสร้างความสับสน และทำให้ประสิทธิภาพ Model ลดลง งานวิจัยในอดีตใช้ Late Fusion หรือ Average Fusion ซึ่งทำให้ประสิทธิภาพลดลงเมื่อเจอปัญหานี้ เพราะแต่ละเฟรมได้ Action Classes ที่หลากหลาย เช่น จากวิดีโอเดียวกัน เฟรม A ได้ คลาสการวิ่ง เฟรม B ได้ คลาสการยิงบาสเก็ตบอล จากปัญหาข้างต้น Rohit Girdhar นำเสนอ ActionVLAD ใน พ.ศ. 2560 ซึ่งเป็นเลเยอร์รวบรวม (Aggregation Layer) ทำหน้าที่เก็บรวบรวม Spatial-Temporal Features แต่ละเฟรม ได้จากทั้งวิดีโอ นอกจากนี้ Girdhar ได้ทดลองการเก็บรวบรวม ActionVLAD ที่ Layer ต่างกัน และการ Fusion Technique แต่ละประเภท เช่น Concatenate Fusion Late Fusion และ Early Fusion ผลการทดลองพบว่า การรวบรวม Features จาก Last Convolution และ Late Fusion ให้ประสิทธิภาพที่ดีที่สุด เทคนิค ActionVLAD ทำงานคล้ายกับ Convolution Layer สามารถนำไปใช้กับ Architecture อื่นได้ และเป็น End-to-end Training

ในขณะเดียวกัน Carreiray และ Zisserman [10] นำเสนอวิธีใหม่ในการทำการรู้จักการกระทำ ใน พ.ศ. 2560 โดย Joao Carreiray มองเห็นว่าข้อมูลโมเดลที่ผ่านการ Training บน Dataset อื่น (Pretrain) สามารถช่วยให้ Model มีประสิทธิภาพสูงขึ้นได้ แต่ว่า Architecture ที่ใช้คอนโวลูชัน 3D ใช้เวลาในการ Training นาน ทำให้ไม่มี Pretrain Model ดังนั้น Joao จึงนำเสนอวิธีในการนำ

ข้อมูล Pretrain Model ที่ Train โดยใช้คอนโวลูชัน 2D มาใช้กับคอนโวลูชัน 3D ด้วยเทคนิคขยาย (Inflated) Filters และเรียกสถาปัตยกรรมนี้ว่า Two-Stream Inflated 3D ConvNet หรือ I3D

งานวิจัยในกลุ่มนี้ใช้ภาพการเคลื่อนไหวเป็นตัวแทนของ Temporal Features ซึ่งทำให้ประสบปัญหา Computation Cost สูงและไม่สามารถนำไปใช้แบบเวลาจริง (Real Time) ได้ ถ้าชุดข้อมูลมีขนาดใหญ่ จากปัญหาข้างต้น Zhu และคณะ [11] นำเสนอ MotionNet ใน พ.ศ. 2561 เพื่อแก้ปัญหา Computation Cost โดยสร้าง Deep Learning Model สำหรับสร้างภาพการเคลื่อนไหวขึ้นมาใช้เป็น Temporal Features สำหรับการรู้จักการกระทำ โดยเริ่มต้น Training Motionnet สำหรับสร้างภาพการเคลื่อนไหว โดย อินพุต เป็นรูปภาพ 2 เฟรม ติดกัน ทำนายภาพการเคลื่อนไหวออกมา โดย Motionnet สามารถทำงานได้รวดเร็ว และไม่ต้องเก็บข้อมูลภาพในคอมพิวเตอร์ทำให้ไม่เสียทรัพยากร รวมถึงสามารถทำงานได้กับสถาปัตยกรรมที่หลากหลาย

ในปีเดียวกันกับ Motionnet นักวิจัยอีกกลุ่มได้ศึกษาวิธีแก้ปัญหา Computation Cost ของการเคลื่อนไหว ซึ่ง Gao และคณะ ได้มองว่าการเคลื่อนไหว เป็น Features ที่สำคัญในการแยกความแตกต่างของการกระทำและการสร้างภาพการเคลื่อนไหวขึ้นใช้เวลาและ Computation Cost ดังนั้น Gao และคณะ [12] ได้พยายามแก้ไขปัญหานี้โดย สร้าง Model ที่ใช้ Encoder-decoder Architecture ในการสร้างภาพการเคลื่อนไหวขึ้นมา โดยเรียนรู้จากวิดีโอที่เกี่ยวกับการเคลื่อนไหว และใช้ Temporal Model นั้น ในการสร้างภาพการเคลื่อนไหวขึ้นมาแทนการใช้ภาพการเคลื่อนไหวจริง ผลลัพธ์ที่ได้ทำให้การ Training สามารถทำได้รวดเร็วมากยิ่งขึ้น และในกรณีที่การกระทำบนข้อมูลที่ไม่เคยพบใน Train Dataset (Unseen Dataset) Temporal Model ยังสามารถให้ประสิทธิภาพสูงอยู่ นอกจากนี้สามารถนำมาใช้ร่วมกับสถาปัตยกรรมอื่นได้

จากงานวิจัยที่ผ่านมาในกลุ่ม Multiple Stream พบว่า Two Stream ใช้การ Train ข้อมูล 2 แบบ ได้แก่ RGB สำหรับ Spatial และการเคลื่อนไหวสำหรับ Temporal และได้ประสิทธิภาพที่สูง แต่มีข้อจำกัดคือ Computational Cost สูง จากการสร้างภาพการเคลื่อนไหว ซึ่งทำให้มีปัญหาไม่สามารถนำไปใช้แบบเวลาจริงได้

2.3 สเปซไทม์ (Space-time)

งานวิจัยในกลุ่มที่ 2 Space-time เป็นงานวิจัยที่ใช้คอนโวลูชัน 3D ซึ่งเป็นเทคนิคที่พัฒนามาจากคอนโวลูชัน 2D โดยสามารถสกัดคุณลักษณะแบบ 3 มิติได้ งานวิจัยในกลุ่มนี้นำคอนโวลูชัน 3D มาใช้ในการแก้ไขปัญหาการรู้จักการกระทำ โดย Tran และคณะ [13] นำเสนอ 3D Convolutional Networks ใน พ.ศ. 2558 โดยนำคอนโวลูชัน 3D มาใช้สกัดคุณลักษณะ Spatial และ Temporal พร้อมกัน และเรียกสถาปัตยกรรมนี้ว่า C3D

ถึงแม้ว่า คอนโวลูชัน 3D สามารถหาข้อมูล Spatial และ Temporal ไปพร้อมกันได้ แต่ยังประสบปัญหา Long Term Dependency เนื่องจาก tran ใช้ Batchsize ขนาดเล็ก ดังนั้น Diba และคณะ [14] ใน พ.ศ. 2560 นำไอเดียจาก GoogleNet ที่มีการหา Feature โดยใช้ Kernel Size ที่มีขนาดต่างกันมาใช้หาข้อมูล Temporal ซึ่งสามารถจับข้อมูลได้ในระยะต้น กลาง ไปจนถึงท้ายข้อมูลได้ รวมถึงประยุกต์ใช้ข้อมูล Pretrain 2D มาใช้เป็นความรู้ตั้งต้น หรือ Weight ให้กับคอนโวลูชัน 3D และเรียกสถาปัตยกรรมนี้ว่า Temporal 3D ConvNet (T3D)

อีกด้านหนึ่ง Zhou และคณะ [15] มีแนวคิดว่าคอนโวลูชัน 3D มีประสิทธิภาพยังไม่เป็นที่พอใจ ถ้าเทียบกับประสิทธิภาพของคอนโวลูชัน 2D ดังนั้น Zhou จึงนำเสนอแนวคิดว่าการผสมข้อมูล คอนโวลูชัน 2D และ 3D เข้าด้วยกันทำให้ได้ประสิทธิภาพสูงขึ้นได้ Zhou เชื่อว่าการใช้คอนโวลูชัน 3D แบบ Layer น้อย และมี Depth เยอะให้ผลดีกว่าคอนโวลูชัน 2D หลาย Layer และ Zhou ตั้งชื่อสถาปัตยกรรมว่า Mixed Convolutional Tube (MiCT) และนำเสนอใน พ.ศ. 2561

จากงานวิจัยที่ผ่านมาในกลุ่ม Space-time พบว่าคอนโวลูชัน 3D สามารถจับข้อมูล Spatial และ Temporal พร้อมกันได้ แต่มีข้อจำกัดคือ Computational Cost สูง

ถ้าใช้ Kernel Size, Depth ที่มีขนาดใหญ่ และในบางกรณีจะติดปัญหา Long Term Dependency

2.4 วิธีที่ใช้เฉพาะเจาะจง (Specific Purpose)

งานวิจัยในกลุ่มที่ 3 วิธีเฉพาะเจาะจง เป็นงานวิจัยที่ใช้วิธีการเฉพาะในการแก้ไขปัญหาการรู้จักการกระทำ เช่น มีการนำ Recurrent Layer เช่น RNN, LSTM หรือ GRU มาใช้รู้จำ Action sequence หรือการสร้างข้อมูล Temporal รูปแบบอื่น นอกเหนือจากการเคลื่อนไหว หรือสนใจในรูปแบบข้อมูลที่ต่างกันของการรู้จักการกระทำ

Donahue และคณะ [16] นำเสนอ Long-term Recurrent Convolutional Networks (LRCN) ใน พ.ศ. 2559 ซึ่งเป็นการนำ Recurrent Neural Network มาใช้ Donahue มีแนวคิดว่า ควร Training แบบ End-to-end โดยใช้คอนโวลูชัน ในการสกัดคุณลักษณะ Spatial และ Recurrent Layer ในการจัดสกัดคุณลักษณะ โดยวิธีการนี้เป็นวิธีที่สามารถนำไปประยุกต์ใช้ได้ง่ายหลายสายงาน เช่น การรู้จักการกระทำ การจับภาพ หรือ Video Description

Wang และคณะ [17] มองว่า Architecture ที่ผ่านมาในอดีตเก็บข้อมูล Temporal หรือ Context ที่ Stage ท้ายใกล้กับ Fully Connect หรือใช้ข้อมูลการเคลื่อนไหวแทนข้อมูล Temporal ซึ่งทำให้มีปัญหา Computation Cost และการเก็บข้อมูล Context ที่ Stage ท้าย ทำให้ข้อมูล Context ที่ได้ ขาดหายไปบางส่วนจากคอนโวลูชัน หรือ Pooling ด้วยเหตุนี้ Wang ได้นำเสนอ Deep Alternative Neural Network (DANN) ใน พ.ศ. 2559

Girdhar และ Ramanan [18] ค้นพบปัญหาของการรู้จักการกระทำในกลุ่มที่นำ Human Key Pose หรือ Object Person มาใช้เป็น Feature ที่ช่วยเพิ่มประสิทธิภาพของการรู้จักการกระทำว่ามีลักษณะความสนใจแบบหนัก (Hard-Code Attention) ซึ่งพิจารณาว่าทุกการกระทำให้ความสำคัญกับ Human part หรือ Object มากไป ซึ่งในบางกรณี เช่น กรณีสามารถแยกความแตกต่างของการกระทำได้จาก Background/Context ได้ เช่น การแยกความต่างของการเตะบอล และการว่ายน้ำ ซึ่งมีความแตกต่างที่ Background อย่างชัดเจน นอกจากนี้การหา Dataset ที่มี Human Part หรือ Bounding Box หาได้ยาก หรือถ้าสร้างขึ้นมาใหม่ก็ต้องใช้ Computation Cost สูง จากปัญหาข้างต้น นำเสนอ Attentional Pooling ใน พ.ศ. 2560 โดยเป็นที่สามารถหา ความสนใจแบบเบา (Soft Attention) จากรูปภาพได้ ซึ่งเป็นเทคนิคที่ทำงานคล้ายกับการทำ Pooling และสามารถต่อเชื่อมเข้ากับสถาปัตยกรรมอื่นได้ นอกจากนี้ยังใช้เวลาในการคำนวนน้อยหรือใกล้เคียงกับสถาปัตยกรรมเดิม

Vahora และ Chauhan [19] ก็เป็นนักวิจัยอีกกลุ่มที่มองปัญหาของการรู้จักการกระทำในรูปแบบใหม่ โดยพิจารณาที่รูปแบบต่างกัน 3 ประเภท ได้แก่ 1) การรู้จำกิจกรรมของมนุษย์เพียงหนึ่งคน (Singular Human Activity Recognition) 2) การรู้จำกิจกรรมแบบกลุ่ม (Group Activity Recognition) และ 3) การวิเคราะห์สิ่งแวดล้อมที่แออัด (Crowded Scene Analysis) ซึ่ง Vahora สนใจไปในการรู้จำกิจกรรมแบบกลุ่ม ซึ่งมีความท้าทายตรงที่จำนวนสมาชิกในกลุ่มที่ทำ Activity ความสัมพันธ์ระหว่างสมาชิกในกลุ่ม ความเคลื่อนไหวของสมาชิกในกลุ่ม และความเคลื่อนไหวของกลุ่มตัวอย่างของการรู้จำกิจกรรมแบบกลุ่ม เช่น การเดินไปพร้อมกันของกลุ่ม การพูดคุยกัน การต่อแถวซื้อของ Vahora นำเสนอ Spatiotemporal Hierarchical Deep Neural Network ใน พ.ศ. 2561 สำหรับใช้รู้จำกิจกรรมแบบกลุ่มโดยใช้เทคนิค ความสัมพันธ์ตามบริบท (Contextual Relationship) เริ่มต้น ตรวจจับหาบุคคลโดยใช้ Bounding Box จากนั้นทำการสกัด Features รายบุคคล โดยใช้ Pretrain ImageNet CNN ต่อมา นำไปเป็น Input ของ Recurrent Network เพื่อหาการรู้จักการกระทำแบบกลุ่ม นอกจากการสกัด Features รายบุคคลแล้ว ยังมีการสกัด Features จากทั้งรูปภาพ เพื่อนำไปพิจารณา Vontext รวมของ Activity

Zhao และคณะ [20] มองเห็นว่าปัญหาในกลุ่ม Space-time จากคอนโวลูชัน 3D ทำให้เกิด Computation Cost สูง เมื่อใช้ Kernel Size ที่มีขนาดใหญ่ และได้มีงานวิจัย

อีกหลายชิ้นที่ประยุกต์ใช้ คอนโวลูชัน 2D หรือ 1D มาช่วยในการหา Temporal Features แต่งานวิจัยเหล่านั้นทำงานบนสมมติฐานที่ว่า Features ที่ต้องการหาต้องไม่เคลื่อนที่ห่างจากตำแหน่งเดิมมาก (Well align) เพื่อเก็บข้อมูลการเคลื่อนไหวที่ตำแหน่งเดียวกันได้ ในขณะที่บางการกระทำมีการเคลื่อนไหวที่คนละตำแหน่งของรูปภาพ เช่น การวิ่งมีการเคลื่อนไหวได้จากทุกทิศทาง เช่น วิ่งจากซ้ายไปขวา จากปัญหาข้างต้น Zhao นำเสนอ คอนโวลูชันแบบทราเจกทรอลี (Trajectory Convolution) ใน พ.ศ. 2561 ทำหน้าที่ สกัด Feature ในจุด Physical เดียวกันในทุกจุด ไม่ใช่ที่พิกเซลเดิม ทำให้สามารถหา Features ที่ Track ตามการเคลื่อนไหวได้

Wu และคณะ [21] ได้นำเสนอวิธีทำการรู้จักการกระทำใน พ.ศ. 2561 โดยสกัด Features จากวิดีโอที่ถูกบีบอัด (Compressed Video) จาก Algorithm MPEG-4, H.264 หรือ HEVC แทนการหาจากวิดีโอปกติ เพราะ Wu มองเห็นปัญหาของการใช้วิดีโอปกติที่ประกอบด้วยข้อมูลซ้ำกันจำนวนมหาศาล ทำให้ใช้เวลา Training นาน และรูปภาพอย่างเดียวสกัดข้อมูลการเคลื่อนไหวได้ยากต้องใช้วิธีอื่นมาช่วยในการหาข้อมูลการเคลื่อนไหว เช่น RNN, คอนโวลูชัน 3D หรือการเคลื่อนไหว ซึ่งสามารถใช้วิดีโอที่ถูกบีบอัดลดปัญหาลงได้ เพราะข้อมูลถูกบีบอัดจะถูกลดขนาดข้อมูลที่ซ้ำกันลงอย่างมหาศาล เนื่องจากลดขนาดข้อมูลลงโดยเก็บเฉพาะเฟรมภาพที่จำเป็น และสร้างเฟรมภาพสำคัญจากเวกเตอร์การเคลื่อนที่ (Motion Vector) และความผิดพลาดแบบเรสสิดวล (Residual Error) ซึ่งเก็บการเคลื่อนที่ของวัตถุในวิดีโอ ประโยชน์คือเก็บเฉพาะข้อมูลการเคลื่อนไหว ไม่สนใจข้อมูลลักษณะเชิงพื้นที่ เช่น ในกรณีที่มี 2 คน ใส่เสื้อผ้าสี รูปทรงต่างกัน ทำการกระทำเดียวกัน จะได้ Motion Signals ที่เหมือนกันหรือคล้ายกัน ซึ่งเป็น Features สำคัญของการทำการรู้จักการกระทำ นอกจากนี้การใช้วิดีโอที่ถูกบีบอัดยังทำให้สามารถ Training ได้ในเวลาที่รวดเร็วและได้ประสิทธิภาพสูงอีกด้วย

Choutas และคณะ [22] มองว่าข้อมูลโครงร่างกระดูกของมนุษย์ (Human Skeleton) เป็น Features สำคัญที่สามารถช่วยแยกความแตกต่างของการกระทำได้ แต่ว่าข้อมูล Skeleton หาได้ยาก มีจำนวนน้อย ดังนั้น Choutas นำเสนอ Features ได้แก่ Potion ซึ่งมาจาก Pose และ Motion ใน พ.ศ. 2561 ซึ่งเป็นวิธีการหาข้อต่อโครงกระดูกมนุษย์ (Human pose Feature) จาก Library Openpose โดยวิธีนี้สามารถจับ Long Term Dependency เริ่มต้นโดยรัน Pose Estimator เพื่อสกัด Human Joints Heatmap ในแต่ละ Frame แต่ละ Pixel ใน Heatmap บอกความน่าจะเป็นที่เป็น Joints นั้น จากนั้นรวบรวมเข้าด้วยกันเพื่อเป็น Potion Feature ด้วยวิธีนี้ทำให้สามารถใช้แค่ Layer จำนวนและเวลาเล็กน้อย นำไปใช้กับสถาปัตยกรรมอื่นได้

ต่อมาจากปัญหาเดียวกับ Potion ที่มองว่าข้อมูลโครงกระดูกมนุษย์สามารถนำไปใช้เป็นคุณลักษณะสำคัญ ได้ Li และคณะ [23] ได้มองเห็นปัญหาว่า การแยกสกัด Features แยกแต่ละข้อต่อ (Joint) ทำให้ขาดข้อมูลความสัมพันธ์ของแต่ละข้อต่อ (Joint Relationship) ซึ่งมีประโยชน์และเป็น Features ที่สำคัญ เช่น ความสัมพันธ์ของมือและเท้าเวลาเดิน โยน หยิบสิ่งของ จากปัญหาข้างต้น Li ได้นำเสนอ Actional-structural Graph Convolution Network (AS-GCN) ใน พ.ศ. 2562 ที่ใช้ Graph Convolution Network ในการสกัดคุณลักษณะและทำการรู้จักการกระทำ

จากปัญหาเดียวกันกับ MotionNet ในกลุ่ม Space-time เรื่อง Computation cost ของภาพการเคลื่อนไหว ทำให้ Crasto และคณะ [24] นำเสนอ Motion-Augmented RGB Stream (MARS) ใน พ.ศ. 2562 ซึ่งเป็น Architecture ที่สามารถใช้ข้อมูลลักษณะเชิงพื้นที่ และข้อมูลการเคลื่อนไหวจากข้อมูลภาพ RGB เท่านั้น โดยใช้เทคนิค Transfer Learning หลักการทำงานคือ Training Flow Stream (Teacher) ด้วยภาพการเคลื่อนไหว ต่อมา Freeze Network ไว้ จากนั้น Training Mars Architecture (Student) โดยใช้ข้อมูลจาก Flow Stream ปรับค่า Loss โดยใช้ Categorical Cross Entropy และ MSE ระหว่าง Mars และ Flow Stream ผลการศึกษาพบว่า สามารถทำงานได้รวดเร็วและมีประสิทธิภาพสูง

Martinez และคณะ [25] นำเสนอวิธีใหม่ในการทำการรู้จักการกระทำ ใน พ.ศ. 2562 โดยผู้วิจัยมองเห็นว่าการกระทำบางประเภทต้องใช้รายละเอียดเชิงลึก (Fine Grain

Detail) ในการระบุการกระทำ เช่น การหยิบโทรศัพท์ขึ้นมาใช้และการหยิบแฮมเบอร์เกอร์ขึ้นมากิน จากการกระทำข้างต้นทำให้บางครั้งโมเดลสับสนเวลาเจอรายละเอียดเชิงลึก ผู้วิจัยนำเสนอเทคนิคการใช้ Discriminative Filter Banks ในการระบุ บริเวณที่เป็นความแตกต่างชัดเจนของ 2 Class ที่คล้ายกัน โดยเทคนิคนี้สามารถนำไปใช้ได้ทั้ง 2D และ 3D Convolution Architecture

จากงานวิจัยที่ผ่านมาในกลุ่ม Specific Purpose พบว่ามีการใช้เทคนิคที่หลากหลายในการจัดการปัญหา เช่น มีการใช้ Human Joint จาก Openpose หรือมีการพัฒนาเทคนิคอื่นขึ้นมา เช่น Trajectory Convolution, Attention Pooling หรือมองปัญหาในมุมมองใหม่ เช่น Discriminative Filter Banks ที่ออกแบบมาเพื่อจัดการกับรายละเอียดเชิงลึก ซึ่งแต่ละวิธีออกแบบมาเพื่อจัดการปัญหาหลากหลายของการรู้จักการกระทำ

3. สรุป

บทความปริทัศน์นี้รวบรวมข้อมูลงานวิจัยการรู้จำการกระทำมาสรุปและเรียบเรียง โดยเลือกงานวิจัยที่มีโจทย์ปัญหา วิธีการ และผลลัพธ์ที่น่าสนใจมานำเสนอ งานวิจัยในอดีตแบ่งออกเป็น 3 กลุ่ม ได้แก่ 1) Multiple Streams 2) Space-time และ 3) Specific propose งานวิจัยทั้ง 3 กลุ่ม มีข้อดีและด้อย งานวิจัยที่ผ่านมาในกลุ่ม Multiple Stream พบว่า Two Stream ใช้การ Train ข้อมูล 2 แบบ ได้แก่ RGB สำหรับ Spatial และการเคลื่อนไหว สำหรับ Temporal และได้ประสิทธิภาพที่สูง แต่มีข้อจำกัดคือ Computational Cost สูง จากการสร้างภาพการเคลื่อนไหว ซึ่งทำให้มีปัญหาไม่สามารถนำไปใช้แบบเวลาจริงได้ จากการศึกษางานวิจัยในกลุ่ม Space-time พบว่า คอนโวลูชัน 3D สามารถจับข้อมูล Spatial และ Temporal พร้อมกันได้ แต่มีข้อจำกัดคือ Computational Cost สูง ถ้าใช้ Kernel Size, Depth ที่มีขนาดใหญ่ และในบางกรณีจะติดปัญหา Long Term Dependency จากงานวิจัยที่ผ่านมาในกลุ่ม Specific Purpose พบว่า มีการใช้เทคนิคที่หลากหลายในการจัดการปัญหา เช่น มีการใช้ Human Joint จาก Openpose หรือมีการพัฒนาเทคนิคอื่นขึ้นมา เช่น Trajectory Convolution, Attention Pooling หรือมองปัญหาในมุมมองใหม่ เช่น Discriminative Filter Banks ที่ออกแบบมาเพื่อจัดการกับรายละเอียดเชิงลึก ซึ่งแต่ละวิธีออกแบบมาเพื่อจัดการปัญหาหลากหลายของการรู้จักการกระทำ

เอกสารอ้างอิง

[1] I. E. Olatunji and C. H. Cheng, “Video analytics for visual surveillance and applications: An overview and survey,” in *Learning and Analytics in Intelligent Systems*. Springer, Cham, 2019.

[2] M. Ravinder, T. V. Gopal, and T. V. N. Rao, “Video indexing and retrieval - applications and challenges,” *Oriental Journal of Computer Science & Technology*, vol. 3, pp. 125–137, 2010.

[3] S. Akihiko, K. Kiichi, K. Masahiro, H. Shoichi, N. Masayuki, and S. Makoto, “Entertainment applications of human-scale virtual reality systems,” in *Advances in Multimedia Information Processing - PCM 2004*. Springer, Berlin, Heidelberg, 2004.

[4] J. Prakash and L. M. Nithya, “A survey on semi-supervised learning techniques,” *International Journal of Emerging Trends & Technology in Computer Science*, vol. 8, no. 1, pp. 25–29, 2014.

[5] M. Luca, W. Xinyi, and P. Anastasia, “Mind the gap: Developments in autonomous driving research and the sustainability challenge,” *Journal of Cleaner Production*, vol. 275, 2020.

[6] S. Karen and Z. Andrew. (2014, November). *Two-stream convolutional networks for action recognition in videos*. [Online]. Available: https://arxiv.org/abs/1406.2199v2

[7] W. Limin, X. Yuanjun, W. Zhe, Q. Yu, L. Dahua,

T. Xiaoou, and V. G. Luc, "Temporal segment networks: Towards good practices for deep action recognition," in *European Conference on Computer Vision*, Springer, Cham, 2016.

[8] F. Christoph, P. Axel, and Z. Andrew, "Convolutional two-stream network fusion for video action recognition," presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[9] G. Rohit, R. Deva, M. Harikrishna, and S. Josef, "Action VLAD: Learning spatio-temporal aggregation for action classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 971–980.

[10] C. Joao and Z. Andrew, "Quo vadis, action recognition? a new model and the kinetics dataset," presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[11] Z. Yi, L. Zhenzhong, N. Shawn, and G. H. Alexander, "Hidden two-stream convolutional networks for action recognition," in *Computer Vision – ACCV 2018*, Springer, Cham, 2018.

[12] R. Gao, B. Xiong, and K. Grauman, "Im 2Flow: Motion hallucination from static images for action recognition," presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[13] T. Du, B. Lubomir, F. Rob, T. Lorenzo, and P. Manohar, "Learning spatiotemporal features with 3D convolutional networks," presented at the 2015 IEEE International Conference on Computer Vision (ICCV), 2015.

[14] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. V. Gool. (2017, November). *Temporal 3D ConvNets: New architecture and transfer learning for video lassification* [Online]. Available: https://arxiv.org/abs/1711.08200

[15] Y. Zhou, X. Sun, Z.-J. Zha, and W. Zeng, "MiCT: Mixed 3D/2D convolutional tube for human action recognition," presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[16] D. Jeff, A. H. Lisa, R. Marcus, V. Subhashini, G. Sergio, S. Kate, and D. Trevor, "Long-term recurrent convolutional networks for visual recognition and description," presented at the IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[17] J. Wang, W. Wang, X. Chen, R. Wang, and W. Gao, "Deep alternative neural network: Exploring contexts as early as possible for action recognition," presented at the Advances in Neural Information Processing Systems, 2016.

[18] G. Rohit and R. Deva, "Attentional pooling for action recognition," presented at the Conference on Neural Information Processing Systems, 2017.

[19] V. Safvan and C. C. Narendrasinh, "Deep neural network model for group activity recognition using contextual relationship," *Engineering Science and Technology an International Journal*, vol. 22, no. 1, pp. 47–54, 2019.

[20] Z. Yue, X. Yuanjun, and L. Dahua, "Trajectory convolution for action recognition," presented at the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada, 2018.

[21] C. Wu, M. Zaheer, H. Hu, R. Manmatha, A. J. Smola, and P. Krähenbühl, "Compressed Video Action Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6026–6035.

[22] C. Vasileios, W. Philippe, R. Jerome, and S. Cordelia, "PoTion: Pose motion representation for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7024–7033.

[23] L. Maosen, C. Siheng, C. Xu, Z. Ya, W. Yanfeng, and T. Qi, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3595–3603.

[24] C. Nieves, W. Philippe, A. Karteek, and S. Cordelia, "MARS: Motion-augmented rgb stream for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7882–7891.

[25] M. M. Brais, M. Davide, X. Yuanjun, and T. Joseph, "Action recognition with spatial-temporal discriminative filter banks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 5482–5491.