การวิเคราะห์ท่าทางของมนุษย์โดยใช้ข้อมูลสีและความลึกจากหลายมุมมอง

Human Action Analysis from Multi-View using RGB-D Information

พงศกร เจริญเนตรกุล

Pongsagorn Chalearnnetkul

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา
ปรัชญาดุษฎีบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์
มหาวิทยาลัยสงขลานครินทร์

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Engineering
Prince of Songkla University
2561

การวิเคราะห์ท่าทางของมนุษย์โดยใช้ข้อมูลสีและความลึกจากหลายมุมมอง

Human Action Analysis from Multi-View using RGB-D Information

พงศกร เจริญเนตรกุล

Pongsagorn Chalearnnetkul

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา
ปรัชญาดุษฎีบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์
มหาวิทยาลัยสงขลานครินทร์

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Engineering
Prince of Songkla University
2561

ชื่อวิทยานิพนธ์ การวิเคราะห์ท่าทางของมนุษย์โดยใช้ข้อมูลสีและความลึกจากหลายมุมมอง
ผู้เขียน นายพงศกร เจริญเนตรกุล
สาขาวิชา วิศวกรรมคอมพิวเตอร์

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	**คณะกรรมการสอบ**
.....	.....ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.นิคม สุวรรณวร)	(รองศาสตราจารย์ ดร.วัฒนพงศ์ เกิดทองมี)
	.....กรรมการ
	(ผู้ช่วยศาสตราจารย์ ดร.นิคม สุวรรณวร)
	.....กรรมการ
	(รองศาสตราจารย์ ดร.มนตรี กาญจนะเดชะ)
	.....กรรมการ
	(รองศาสตราจารย์ ดร.พรชัย พฤกษ์ภัทรานนต์)
	.....กรรมการ
	(ดร.อนันท์ ชกสุริวงค์)

บัณฑิตวิทยาลัย มหาวิทยาลัยสงขลานครินทร์ อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษา ตามหลักสูตรปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์

.....

(ศาสตราจารย์ ดร.ดำรงศักดิ์ ฟ้ารุ่งสาง)
คณบดีบัณฑิตวิทยาลัย

ขอรับรองว่า ผลงานวิจัยนี้มาจากการศึกษาวิจัยของนักศึกษาเอง และได้แสดงความขอบคุณบุคคลที่มีส่วนช่วยเหลือแล้ว

ลงชื่อ.....
(ผู้ช่วยศาสตราจารย์ ดร.นิคม สุวรรณวร)
อาจารย์ที่ปรึกษาวิทยานิพนธ์

ลงชื่อ.....
(นายพงศกร เจริญเนตรกุล)
นักศึกษา

ข้าพเจ้าขอรับรองว่า ผลงานวิจัยนี้ไม่เคยเป็นส่วนหนึ่งในการอนุมัติปริญญาในระดับใดมาก่อน และไม่ได้ถูกใช้ในการยื่นขออนุมัติปริญญาในขณะนี้

ลงชื่อ.....
(นายพงศกร เจริญเนตรกุล)
นักศึกษา

ชื่อวิทยานิพนธ์ การวิเคราะห์ท่าทางของมนุษย์โดยใช้ข้อมูลสีและความลึกจากหลายมุมมอง
ผู้เขียน นายพงศกร เจริญเนตรกุล
สาขาวิชา วิศวกรรมคอมพิวเตอร์
ปีการศึกษา 2561

บทคัดย่อ

งานวิจัยฉบับนี้ได้ประยุกต์ใช้เทคนิคด้านการประมวลผลภาพและคอมพิวเตอร์วิทัศน์ เพื่อใช้ในการสร้างแบบจำลองโครงสร้างมนุษย์เพื่อใช้การรู้จำท่าทางพื้นฐานของมนุษย์ รวมไปถึงการติดตามและจดจำตัวบุคคล การตรวจจับเหตุการณ์ที่น่าสนใจ ในกรณีศึกษาของการล้ม การโบกมือเพื่อขอความช่วยเหลือ และการกระโดด จากภาพสีและความลึกหลายมุมมอง ณ บริเวณเดียวกัน เพื่อนำไปใช้ในการในการเฝ้าระวังด้านงานส่งเสริมสุขภาพและระบบรักษาความปลอดภัย โดยในงานส่วนของการรู้จำท่าทางจะมีการใช้การฟิวชันข้อมูลระดับสูง และระดับฟีเจอร์ ซึ่งในการฟิวชันข้อมูลระดับสูง จะนำท่าทางจากระบบการรู้จำท่าทางจากมุมมองเดี่ยวหลาย ๆ มุมมองมาเพื่อตัดสินใจ ชั่งน้ำหนัก วัดค่าความน่าเชื่อถือของคำตอบจากการทำนายคำตอบในมุมมองเดี่ยวต่าง ๆ และหาท่าทางที่มีความน่าเชื่อถือที่สุดมาเป็นคำตอบของระบบโดยรวม โดยได้เพิ่มความถูกต้องของการรู้จำจากมุมมองเดี่ยวมากที่สุดที่ 98.17% และได้เพิ่มความถูกต้องจากการเฉลี่ยทุกมุมมองและท่าทางที่ 16.66% (แต้มเปอร์เซ็นต์) จากการเปรียบเทียบกับมุมมองเดี่ยว ส่วนการฟิวชันข้อมูลระดับฟีเจอร์นั้น ผู้วิจัยได้นำเสนอวิธีการที่จะสร้างแบบจำลองฟีเจอร์แบบเลเยอร์ (Layer Feature Model) ซึ่งทำให้สามารถฟิวชันฟีเจอร์ของข้อมูลความลึกจากหลายมุมมอง โดยผลการทดลอง ได้ทดสอบในชุดข้อมูล Northwestern UCLA มีความแม่นยำเฉลี่ย 86.40% ชุดข้อมูล i3DPost ที่มีความแม่นยำเฉลี่ยที่ 93.00% และความแม่นยำเฉลี่ย 99.31% ในชุดข้อมูล PSU ในส่วนของการติดตามและจดจำตัวบุคคลนั้น จะใช้การวิเคราะห์ข้อมูลตำแหน่งและสี ซึ่งเป็นข้อมูลเบื้องต้นที่เด่นชัดสามารถนำมาใช้ในการแยกแยะแต่ละบุคคลเพื่อติดตามทั้งในกล้องเดียวกันและระหว่างกล้อง รวมไปถึงใช้ในการจดจำตัวบุคคลในเบื้องต้น ซึ่งการติดตามและจดจำตัวบุคคลมีความแม่นยำในกรณีที่เข้าไปในระบบครั้งละหนึ่งคนที่ 92.87% และกรณีที่เข้าไปในระบบครั้งละสองคนที่ 85.50% ส่วนในการตรวจจับเหตุการณ์ที่น่าสนใจในกรณีศึกษาของการล้ม มีความแม่นยำที่ 90.65% ตามการรู้จำท่าทางในท่านอน การโบกมือเพื่อขอความช่วยเหลือมีอัตราความแม่นยำการตรวจจับได้โดยเฉลี่ยทั้งหมด 92.96% และการกระโดดมีค่าความไวในการตรวจจับโดยเฉลี่ยเท่ากับ 94.44% และค่าความจำเพาะโดยเฉลี่ยเท่ากับ 99.31%

คำสำคัญ : คอมพิวเตอร์วิทัศน์, รู้จำ, ท่าทางมนุษย์, หลายมุมมอง, เลเยอร์, สีและความลึก

Thesis Title Human Action Analysis from Multi-View using RGB-D Information
Author Mr. Pongsagorn Chalearnnetkul
Major Program Computer Engineering
Academic Year 2018

ABSTRACT

This research applied image processing and computer vision techniques to contribute three essential perspectives; modeling profile-based human action recognition, people tracking and re-identification, interesting event detection. The interesting event detection consists of falling, hand-waving for asking help, and jumping. This approach is based on overlapped area of interest using multi-view RGB-D. All functions are purposed for health care promotions and surveillance system applications. The action recognition consists of high level and feature level fusion. In high level, the results of action in single-view system are fused for making decision using empirical analysis to weight the most realizable result to be a result of multi-view system. The maximum improvement for some action is up to 97.70% and overall result increases to 16.66% (percentage point) when compared with single-view action recognition. Another in feature level fusion, we proposed a method to build Layer Feature Model that allows to fuse features of depth from multi-view. The experimental results of fusion model are 86.40% in NW-UCLA dataset, 93.00% in i3DPost dataset, and 99.31% in PSU dataset. In addition, we introduce people tracking and people re-identification by using analysis of position and color descriptor. The position and color descriptor are clearly attributes for both tracking in a single-view and matching those views. Moreover, the color descriptor is also used for supporting cursory people re-identification. The precisions of people re-identification are 92.87% in single person entering and 85.50% when 2-person simultaneously entering. In the interesting event detection, falling detection resulted in the average precision of 90.65% that derived from precision of lying. The average of hand-waving detection precision is 92.96%. The jumping detection has sensitivity rate 92.96% and specificity rate 99.31%.

KEYWORDS : Computer vision, Recognition, Human action, Multi-view, Layer, RGB-D

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ได้ดำเนินการวิจัยและจัดทำวิทยานิพนธ์ได้สำเร็จได้ลุล่วงด้วยดีโดยการสนับสนุนจากบุคคลต่าง ๆหลายฝ่าย ทางผู้วิจัยขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.นิคม สุวรรณวร ที่ปรึกษาวิทยานิพนธ์เป็นอย่างสูง ที่ได้เมตตาให้คำปรึกษาในการทำวิจัยและวิทยานิพนธ์ฉบับนี้ ขอขอบพระคุณคณาจารย์ทุกท่านที่ได้มอบความรู้ต่าง ๆที่ใช้ในการทำวิทยานิพนธ์ฉบับนี้ ขอขอบพระคุณกรรมการสอบความก้าวหน้าวิทยานิพนธ์ทุกท่านที่ได้ติดตามและให้คำแนะนำต่าง ๆซึ่งเป็นประโยชน์เป็นอย่างมากในการทำวิจัย และขอขอบพระคุณบุคคลากรในภาควิชาวิศวกรรมคอมพิวเตอร์ทุกท่านที่ให้ความช่วยเหลือในระหว่างการทำวิทยานิพนธ์

ขอกราบขอบพระคุณประธานกรรมการและกรรมการสอบป้องกันวิทยานิพนธ์ทุกท่าน ที่ได้มีส่วนช่วยการในตรวจทานและให้คำแนะนำในการแก้ไขวิทยานิพนธ์ฉบับนี้ให้ดียิ่งขึ้น

ขอกราบขอบพระคุณ บิดา มารดาและญาติพี่น้องทุกคน ซึ่งเป็นผู้มีพระคุณสูงสุดที่ให้กำลังใจและให้การสนับสนุนด้านการศึกษาด้วยดีมาโดยตลอด

ขอขอบพระคุณนักศึกษาในภาควิชาวิศวกรรมคอมพิวเตอร์ที่ได้มีส่วนช่วยในการเก็บวีดีโอชุดข้อมูลสำหรับใช้ทดสอบระบบต่าง ๆที่อยู่ในวิทยานิพนธ์นี้ รวมถึงการให้คำปรึกษาชี้แนะและกำลังใจจากเพื่อนนักศึกษาทุกคนที่มีส่วนช่วยทำให้วิทยานิพนธ์นี้สำเร็จลุล่วงไปด้วยดี

ขอขอบพระคุณบัณฑิตวิทยาลัยที่ได้อนุมัติทุนอุดหนุนการวิจัยเพื่อวิทยานิพนธ์ประจำปี 2558 ซึ่งมีส่วนช่วยเป็นอย่างมากในการจัดหาเครื่องมือ วัสดุ อุปกรณ์ ต่าง ๆ ที่มีความจำเป็นในการทำวิทยานิพนธ์ครั้งนี้

ขอขอบพระคุณศูนย์ความเป็นเลิศด้านชีววิทยาศาสตร์ ที่มีส่วนช่วยในการสนับสนุนงานวิจัยด้านระบบเฝ้าระวังผู้สูงอายุ ซึ่งเป็นงานส่วนหนึ่งของวิทยานิพนธ์ฉบับนี้

ขอกราบขอบพระคุณอาจารย์วิวัฒน์ สุทธิวิภากร ที่มีช่วยในการแก้ไขปรับปรุงต้นฉบับวารสารวิชาการที่เป็นส่วนหนึ่งของวิทยานิพนธ์ฉบับนี้ ซึ่งทำให้ได้รับการตอบรับตีพิมพ์ในวารสารระดับนานาชาติ

สุดท้ายนี้ขอขอบพระคุณมหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย ที่ได้อนุมัติทุนการศึกษาตามโครงการผลิตและพัฒนาบุคลากรมหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัยสำหรับบุคคลทั่วไป ซึ่งเป็นกำลังสนับสนุนของข้าพเจ้าที่สำคัญยิ่งในการศึกษาและทำวิทยานิพนธ์ฉบับนี้

พงศกร เจริญเนตรกุล

สารบัญ

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญ (ต่อ)

รายการตาราง

รายการตาราง (ต่อ)

รายการตาราง (ต่อ)

รายการภาพประกอบ

รายการภาพประกอบ (ต่อ)

รายการภาพประกอบ (ต่อ)

รายการภาพประกอบ (ต่อ)

รายการภาพประกอบ (ต่อ)

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของงานวิจัย

ในอดีตและปัจจุบัน (2561) ระบบอัจฉริยะต่าง ๆถูกวิจัยและพัฒนาอย่างต่อเนื่อง ซึ่งต่อไประบบเหล่านี้จะเข้ามามีบทบาทต่าง ๆในชีวิตประจำวันของมนุษย์มากยิ่งขึ้น และในปัจจุบันการรักษาความปลอดภัยโดยการบันทึกภาพวิดีโอจากกล้องวงจรปิดอย่างเดียวนั้น ไม่สามารถตอบโจทย์ในการป้องกันหรือเฝ้าระวังและเตือนได้อย่างทันท่วงที จึงต้องมีการให้บุคคลไปคอยเฝ้าสังเกตการณ์ ซึ่งก็อาจจะมีข้อจำกัดด้านความน่าเชื่อถือในการทำงานเป็นระยะเวลานานของมนุษย์ ระบบอัจฉริยะจึงถูกพัฒนาขึ้นเพื่อนำมาประยุกต์ใช้งานเพื่อวิเคราะห์ ติดตาม ตรวจสอบ และแจ้งเตือนแทนมนุษย์ เมื่อเกิดเหตุการณ์ที่ผิดปกติ โดยระบบอัจฉริยะที่ผู้วิจัยสนใจพัฒนาขึ้นมาคือ ระบบการรู้จำท่าทางพื้นฐานของมนุษย์ (Profiled-based action) ประกอบด้วย ยืน/เดิน นั่ง นอน และก้ม ซึ่งเป็นหน่วยย่อยของการรู้จำท่าทางเป็นสิ่งที่ท้าทายในงานด้านคอมพิวเตอร์วิทัศน์ (Computer Vision) ที่จะนำไปสู่การรู้จำการเข้าใจท่าทางที่ซับซ้อนกว่า กิจกรรม และพฤติกรรมของมนุษย์ที่ประกอบไปด้วยหลายท่าทางที่ต่อเนื่องกันไป ซึ่งจะรวมกันเป็นกิจกรรมต่าง ๆ ตัวอย่างเช่น การนอนหลับจะเกี่ยวเนื่องกับท่าทางการยืน การเดิน การนั่ง และการนอน อีกตัวอย่างหนึ่ง คือการล้ม จะประกอบไปด้วย ท่าทางการยืน การเดิน และการนอน ซึ่งได้มีงานวิจัยประยุกต์ท่าทางพื้นฐานของมนุษย์ไปรู้กิจกรรมและพฤติกรรมที่ซับซ้อนกว่า เช่น การรู้จำกิจกรรมจากท่าทางพื้นฐาน โดยใช้ Statistical Model [1] และ Graph Similarity Measurement [2] โดยรู้จำเป็นกิจกรรมดังต่อไปนี้ การเดินสำรวจ การพักผ่อนที่ผิดปกติ การมองหาบางอย่าง การพักผ่อนตามปกติ การเดินผ่านไป อีกตัวอย่างงานวิจัยหนึ่งเป็นการประยุกต์การรู้จำกิจกรรมจากการวิเคราะห์ท่าทางพื้นฐาน วัตถุ และ สถานที่ [3] โดยมีตัวอย่างกิจกรรมเช่น เดินภายในห้อง ทำงานบนโต๊ะทำงาน เปิดตู้เย็น เป็นต้น

นอกจากนี้ผู้วิจัยได้นำเสนอระบบการติดตามและจดจำตัวบุคคลที่สามารถติดตามและจดจำตัวบุคคลที่เข้าออกในสถานที่หนึ่ง ๆได้ และระบบสุดท้ายที่ผู้วิจัยได้นำเสนอคือ การวิเคราะห์เหตุการณ์ที่น่าสนใจ ได้แก่ การล้มตัวลงนอน การโบกมือเพื่อขอความช่วยเหลือ และการกระโดด โดยสิ่งเหล่านี้สามารถนำไปประยุกต์ใช้กับระบบรักษาความปลอดภัย รวมถึงการดูแลด้านสุขภาพของบุคคล โดยเฉพาะอย่างยิ่งในการเฝ้าระวังผู้สูงอายุที่อยู่บ้านคนเดียว ซึ่งเหมาะกับสถานการณ์ในปัจจุบันที่กำลังเข้าสู่สังคมผู้สูงอายุ

โดยจากการศึกษามานั้น ในงานวิจัยที่ทำการวิเคราะห์เฝ้าระวังโดยกล้องตัวเดียวจะมีผลลัพธ์โดยรวมที่ดีพอควร แต่ในการตรวจจับพื้นที่ที่มีบริเวณกว้าง การใช้กล้องตัวเดียวอาจจะทำให้สูญเสียความสามารถบางอย่างในการตรวจจับท่าทางหรือกิจกรรมของมนุษย์ เนื่องจากมุมมองที่เปลี่ยนแปลงไป การบดบังทั้งจากสิ่งกีดขวางและบุคคล การสูญเสียข้อมูลจากตัวกล้องเอง

ดังนั้น การใช้กล้องหลายตัวจึงได้ถูกนำเสนอเพื่อลดข้อจำกัดที่มีอยู่ในการใช้งานกล้องตัวเดียว โดยวิธีการใช้กล้องหลายตัวนี้จะมีข้อด้อยเนื่องจากจำเป็นต้องใช้กล้องหลายตัว ซึ่งจะเพิ่มค่าใช้จ่าย รวมไปถึงความซับซ้อนในการติดตั้ง ขั้นตอนในการสอบเทียบวัดค่าของกล้อง (Camera Calibration) ระหว่างมุมมองและการสร้างแบบจำลองที่ซับซ้อนขึ้น ซึ่งจะใช้เวลาทั้งการติดตั้งและประมวลผลที่มากขึ้น ในงานวิจัยนี้จะพยายามทำให้การนำไปใช้จริงนั้นติดตั้งให้ง่ายที่สุด มีความยืดหยุ่น และง่ายเท่าที่จะเป็นไปได้ ปราศจากการสอบเทียบค่าของกล้อง หรือมีระบบให้ทำอัตโนมัติระหว่างกล้อง โดยในงานวิจัยนี้ใช้แค่เพียงข้อกำหนดการติดตั้งที่กล้องต้องติดตั้งไว้เหนือหัวทำมุม 60 องศา กับ แนวดิ่ง ณ ยอดสุดของเสา 2 เมตร และการทำมุมระหว่างกล้องไปยังบริเวณที่สนใจได้ตั้งแต่ 30 ถึง 135 องศา

อีกปัญหาสำคัญที่ต้องเผชิญภายใต้การใช้การมองเห็นจากกล้องเพื่อรู้จำท่าทางทั้งกล้องตัวเดียวและหลายตัว คือ ความเป็นส่วนตัวของผู้ถูกตรวจจับและเงื่อนไขของสภาวะแสง โดยเทคนิคที่ได้กล่าวมาข้างต้นเกี่ยวข้องกับการใช้กล้องสีและเซ็นเซอร์อื่น ๆ ซึ่งประสบพบเจอกับปัญหาด้านความเป็นส่วนตัวโดยทั้งสิ้น ระบบการตรวจจับท่าทางในพื้นที่ที่เป็นส่วนตัวโดยใช้กล้องสี จะทำให้ผู้ที่ถูกตรวจจับรู้สึกไม่เป็นส่วนตัว เนื่องจากภาพสีจะเผยลักษณะทางกายภาพของมนุษย์ทั้งหมดอีกทั้งมีเงื่อนไขในเรื่องของสภาวะแสงที่เปลี่ยนแปลงอยู่ตลอดเวลา ซึ่งยากต่อการควบคุม ซึ่งกล้องความลึกสามารถช่วยแก้ปัญหาทั้งสองนี้ได้ โดยใช้แค่ภาพความลึกหยาบ ๆ ที่พอเห็นรูปร่างของบุคคลก็เพียงพอสำหรับการรู้จำท่าทาง และข้อมูลความลึก จากกล้อง RGB-D (RGB and Depth) ซึ่งสามารถหลีกเลี่ยงปัญหาการเปลี่ยนแปลงของแสง ที่เป็นปัญหาที่สำคัญในการใช้งานจริงในชีวิตประจำวันตลอดเวลาทั้งกลางวันและกลางคืน โดยการใช้งานภาพความลึกจะถูกนำมาใช้งานในงานวิจัยนี้ด้วยกันกับการใช้กล้องหลายตัวซึ่งได้ภาพหลายมุมมองที่สังเกตการณ์ไปยังบริเวณเดียวกัน และในส่วนของงานวิจัยด้านนี้ยังมีปัญหาที่ต้องการการปรับปรุงก็คือเรื่องความทนทานต่อการเปลี่ยนแปลงมุมมองที่กล้องตรวจจับไปยังมนุษย์ (Perspective Robustness) และความซับซ้อนของแบบจำลอง ซึ่งในงานวิจัยนี้จะมุ่งเน้นไปพัฒนาในส่วนของ ระบบที่ไม่ใช้การสอบเทียบวัดค่าของกล้อง (Free Calibration) กระบวนการฟิวชันข้อมูลที่ทนทานและง่ายต่อการรู้จำท่าทางพื้นฐานจากภาพความลึก ผู้วิจัยได้พัฒนาแบบจำลองการฟิวชันข้อมูลจากเลเยอร์ (Layer Fusion Model) เพื่อที่จะฟิวชันข้อมูลภาพความลึกจากหลายมุมมองและได้ทดสอบข้อมูลของผู้วิจัยจากชุดข้อมูล 3 ชุด คือ Northwestern UCLA และ PSU ที่เป็นชุดข้อมูลท่าทางภาพสีและความลึก หลายมุมมอง จากข้อมูลกล้องหลายตัว รวมไปถึงชุดข้อมูล i3DPost เป็นชุดข้อมูลท่าทางภาพสีจากหลายมุมมอง

ในระบบรู้จำท่าทางจากการมองเห็น (Visual-based Action Recognition) จะมีปัญหาซึ่งเป็นเงื่อนไขในการใช้งานจริงที่จะต้องสามารถรู้จำตัวบุคคลได้จากทุกมองที่กล้องตรวจจับอยู่ รวมไปถึงประมวลผลได้ทันตามการใช้งานแบบตามเวลาจริง (Real-time Processing) และต้องให้ความเป็นส่วนตัวในการดึงข้อมูลภาพจากการมองเห็นมาใช้วิเคราะห์ ดังนั้น ผู้วิจัยจึงได้นำเสนอ

วิธีการที่จะสร้างแบบจำลองฟีเจอร์แบบเลเยอร์ (Layer Feature Model) เพื่อแก้ปัญหาพื้นฐานเหล่านี้ โดยที่อนุญาตให้การฟิวชันฟีเจอร์ของข้อมูลความลึกจากหลายมุมมอง โดยแบบจำลองนี้สามารถให้การรู้จำในหลายมุมมอง โดยมีความซับซ้อนที่น้อย โดยสามารถประมวลผลได้ทันตามเวลาจริงที่ 63 เฟรมต่อวินาที (Intel Core i5 4590 at 3.30GHz with DDR3 8GB) สำหรับท่าทางพื้นฐาน 4 ท่าทาง โดยผลการทดลองได้ทดสอบในชุดข้อมูล Northwestern UCLA ที่ให้ผลลัพธ์มีความแม่นยำเฉลี่ยที่ 86.40% ซึ่งชุดข้อมูล i3DPost ให้ผลความแม่นยำเฉลี่ยที่ 93.00% และ99.31% ในชุดข้อมูล PSU Profile-based action ซึ่งเป็นชุดข้อมูลใหม่ที่จัดทำขึ้นโดยกลุ่มของผู้วิจัย โดยเป็นภาพหลายมุมมองจากข้อมูลกล้องหลายตัว ซึ่งจะให้ข้อมูลท่าทางพื้นฐานของมนุษย์ที่เป็นภาพความลึกและสี ที่จัดทำขึ้นโดยกลุ่มของผู้วิจัย

นอกจากการสร้างแบบจำลองฟีเจอร์แบบเลเยอร์เพื่อการรู้จำท่าทางการฟิวชันฟีเจอร์ในระดับล่างจากหลายมุมมองแล้ว ผู้วิจัยได้พัฒนาการฟิวชันคำตอบจากการรู้จำท่าทางจากมุมมองเดี่ยวของ P.Chawalitsittikul และคณะ [4] โดยแนวความคิดของการฟิวชันข้อมูลในระดับคำตอบได้มาจากการที่สังเกตการณ์รู้จำท่าทางในหลากหลายมุมมองที่ส่องไปยังมนุษย์ ซึ่งวิธีการนี้จะเพิ่มความถูกต้องของการรู้จำโดยสร้างฟังก์ชันวัดค่าความน่าเชื่อถือของคำตอบ (Weighting Function) โดยใช้ความถูกต้องของการรู้จำในมุมมองต่าง ๆที่ได้สังเกตการณ์เป็นเกณฑ์ โดยระบบเพิ่มความถูกต้องของการรู้จำจากมุมมองเดี่ยวมากที่สุดที่ 98.17% และเพิ่มความถูกต้องโดยเฉลี่ยทุกมุมมองและท่าทางที่ 16.66%

โดยอีกส่วนประกอบหนึ่งที่เป็นส่วนสำคัญของระบบ คือส่วนของการติดตามและจดจำตัวบุคคล ซึ่งจะใช้การวิเคราะห์ข้อมูลตำแหน่ง และสี ซึ่งเป็นข้อมูลเบื้องต้นที่เด่นชัดที่สามารถนำมาใช้ในการแยกแยะแต่ละบุคคลเพื่อติดตามทั้งในกล้องเดียวกันและระหว่างกล้อง รวมไปถึงใช้ในการจดจำตัวบุคคลในขั้นต้น ซึ่งการติดตามและจดจำตัวบุคคลมีความแม่นยำในกรณีที่เข้าไปในระบบครั้งละหนึ่งคนที่ 92.87% และกรณีที่เข้าไปในระบบครั้งละสองคนที่ 85.50%

นอกจากนี้งานวิจัยนี้ยังมุ่งเน้นไปที่การตรวจจับเหตุการณ์ต่าง ๆ ที่มีความน่าสนใจ ซึ่งอาจจะเกิดขึ้นได้ในระบบเฝ้าระวังและดูแลด้านสุขภาพ ได้แก่ การตรวจจับการล้มซึ่งทำให้สามารถเตือนให้ผู้ที่เกี่ยวข้องสามารถให้การช่วยเหลือได้ทันท่วงที การตรวจจับการโบกมือเพื่อขอความช่วยเหลือ ซึ่งในบางครั้งผู้ที่ต้องการความช่วยเหลืออาจจะไม่สามารถเคลื่อนที่ไปกดปุ่มที่อุปกรณ์ส่งสัญญาณขอความช่วยเหลือได้ การโบกมือให้ระบบรู้จำจากการมองเห็นจึงเป็นทางเลือกหนึ่งในการส่งสัญญาณขอความช่วยเหลือได้ และส่วนของระบบการตรวจจับการกระโดด ซึ่งอาจจะเกิดได้ในกรณีของการตกใจจากการถูกจี้ปล้น หรือวิ่งหลบบางอย่าง ซึ่งเป็นเหตุการณ์ที่ไม่ปกติสำหรับในพื้นที่ร่มในอาคาร

ในส่วนตรวจจับเหตุการณ์ที่น่าสนใจในกรณีศึกษาของการล้ม ซึ่งเป็นการต่อยอดการรู้จำเหตุการณ์นี้จากการรู้จำท่าทาง โดยที่จะวิเคราะห์การเปลี่ยนแปลงท่าทางอื่น ๆแล้วเป็นท่านอนประกอบกับสถานที่ที่ล้มต้องไม่ใช่บริเวณเตียงนอน โซฟา หรืออื่น ๆ ที่ควรจะนอน ซึ่งจะมีความ

แม่นยำ 90.65% ตามการรู้จำท่าทางในท่านอน และในส่วนของการตรวจจับการโบกมือเพื่อขอความช่วยเหลือ โดยอาศัยการวิเคราะห์โครงร่างของมนุษย์จากภาพฉายในแนวต่าง ๆ มีอัตราความแม่นยำการตรวจจับได้โดยเฉลี่ยทั้งหมด 92.96% และในส่วนของการกระโดด ซึ่งอาศัยการวิเคราะห์การตรวจจับลำดับของจังหวะดีดตัวขึ้นและจังหวะดิ่งลง มีประสิทธิภาพตามค่าความไวในการตรวจจับโดยเฉลี่ยเท่ากับ 94.44% และค่าความจำเพาะโดยเฉลี่ยเท่ากับ 99.31%

1.2 วัตถุประสงค์

1.2.1 เพื่อวิจัยเทคนิคการวิเคราะห์ท่าทางของมนุษย์และการติดตามและรู้จำตัวบุคคลโดยใช้ข้อมูลภาพสีและความลึกจากหลายมุมมอง

1.2.2 เพื่อพัฒนาแบบจำลองการฟิวชันข้อมูลจากหลายมุมมองในการรู้จำท่าทาง

1.3 งานวิจัยที่เกี่ยวข้อง

1.3.1 การรู้จำท่าทางจากเซ็นเซอร์ความเฉื่อยและภาพจากมุมมองเดี่ยว (Inertial-sensor and Singleview-vision Action Recognition)

ตั้งแต่ปี 2010 งานวิจัยด้านการรู้จำท่าทางเพิ่มขึ้นอย่างต่อเนื่อง และมีการนำเสนอในการใช้งานด้านการดูแลผู้ป่วย โดยเฉพาะอย่างยิ่งเพื่อดูแลและเฝ้าระวังผู้สูงอายุ การวิเคราะห์ท่าทางเข้ามามีบทบาทสำคัญในการตรวจสอบเฝ้าระวัง ทั้งพฤติกรรมที่ปกติและผิดปกติในชีวิตประจำวัน ในการใช้งานที่กล่าวมานั้น ความเป็นส่วนตัวและความสะดวกของการใช้งานจะเป็นปัจจัยสำคัญในการเลือกใช้งานเทคโนโลยีที่ต้องพิจารณาอย่างถี่ถ้วน ในไม่กี่ปีที่ผ่านมานี้ (2018) มีวิธีการ 2 วิธีการหลัก ๆ ที่ได้นำเสนอสำหรับการรู้จำท่าทางคือการใช้งานเซ็นเซอร์ และการใช้การมองเห็นจากกล้อง

การรู้จำท่าทางจากเซ็นเซอร์ความเฉื่อย (Inertial Sensor) ได้ถูกใช้กันอย่างกว้างขวาง เนื่องจากมีขนาดที่เล็ก กินพลังงานต่ำ มีค่าใช้จ่ายต่ำ รวมไปถึงการติดกับตัวหรืออยู่ในอุปกรณ์พกพาอื่น ๆ เช่น โทรศัพท์มือถือ นาฬิกาอัจฉริยะ โดยอุปกรณ์เซ็นเซอร์ความเฉื่อยนี้จะใช้เป็นตัวนำทางโดยรวมประกอบไปด้วย การบอกความเคลื่อนไหว แนวการหมุน อย่างเช่น Accelerometers, Gyroscopes เป็นต้น ซึ่งจะให้ข้อมูลผ่านวิถีของการเคลื่อนที่ จุดที่อยู่ ความเร็ว และความเร่งของบุคคลที่กำลังติดตามตัวอยู่ บางงานวิจัยใช้เซ็นเซอร์ที่สามารถสวมใส่ได้ [5-7] โทรศัพท์มือถือ [8-11] และนาฬิกาอัจฉริยะ [12] เพื่อรู้จำท่าทางที่มีความแตกต่างกัน ในบางงานวิจัยได้เน้นไปเฉพาะที่การตรวจจับท่าทางที่ผิดปกติอย่างเช่น การล้ม [13-15] หรือรายงานสถานะทั้งปกติและผิดปกติ [16] ในการที่จะใช้เซ็นเซอร์มาวิเคราะห์ท่าทางที่ซับซ้อน ต้องใช้เซ็นเซอร์ที่มากยิ่งขึ้นไปติดในตำแหน่งต่าง ๆ ของตัวบุคคลจึงเป็นข้อจำกัดในการใช้งานของ

เซ็นเซอร์ความเฉื่อย ซึ่งไม่สะดวกที่จะใช้งานจริง เนื่องจากต้องติดเซ็นเซอร์และอุปกรณ์ ซึ่งจะทำให้เกิดความไม่สะดวกสบายและความรำคาญ

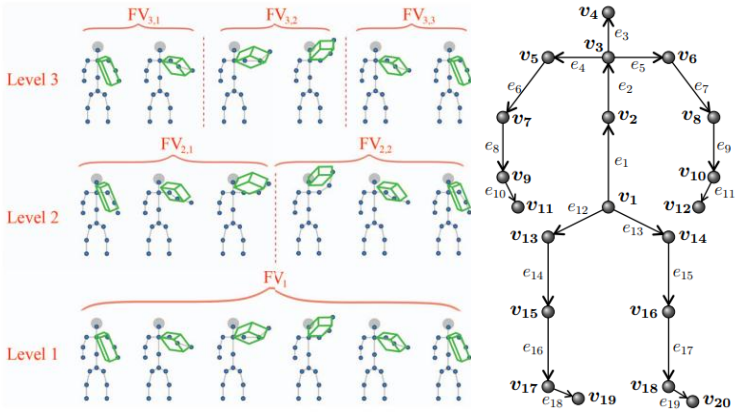
ภาพประกอบที่ 1-1 การติดเซ็นเซอร์และอุปกรณ์กับบุคคลที่ต้องการรู้จำท่าทาง [5]

สำหรับการใช้เทคนิคการมองเห็นเพื่อรู้จำท่าทาง ได้มีหลากหลายงานวิจัย ซึ่งทั้งเน้นไปที่การใช้ข้อมูลจากกล้องตัวเดียวและหลายกล้อง

ในการใช้งานกล้องตัวเดียวมีงานวิจัยหลัก ๆ ซึ่งแบ่งได้ตามการแสดงฟีเจอร์ (Feature Representation) ได้แก่ การแสดงฟีเจอร์เป็นข้อต่อและโครงกระดูก (Joint-based/skeleton-based), การแสดงฟีเจอร์ที่เป็นลักษณะการเคลื่อนไหวหรือการไหล (Motion/flow-based), การแสดงฟีเจอร์ที่เป็นปริมาตรของวัตถุที่สัมพันธ์กับพื้นที่และเวลา (Space-time Volume-based), และ การแสดงฟีเจอร์ในรูปแบบของกริด (Grid-based)

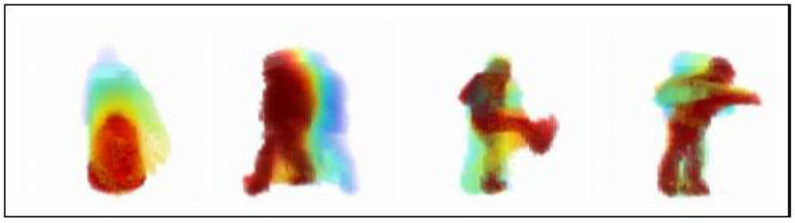
(ก) **การแสดงฟีเจอร์เป็นข้อต่อและโครงกระดูก (Joint-based/skeleton-based)** จะเป็นการกำหนดคุณลักษณะของโครงสร้างพื้นฐานของมนุษย์ โดยใช้ตำแหน่งของข้อต่อหรือโครงชิ้นส่วนหลักต่าง ๆของมนุษย์ในการแยกท่าทาง ตัวอย่างเช่น ข้อมูลของข้อต่อและส่วนต่าง ๆที่มีหลายระดับจากฟีเจอร์โพส (Posture) [17], เวกเตอร์ Fisher โดยใช้ Quads Skeleton [18], ข้อมูลเชิงพื้นที่และเวลา (Spatial-temporal) ของข้อต่อจาก mHOG [19], Lie Vector Space จากโครงร่างแบบกระดูก 3 มิติ (3D Skeleton) [20], การติดตามเส้นทางเคลื่อนที่จากพื้นที่ที่ไม่เปลี่ยนแปลง (Invariance Space Trajectories Tracking) โดยใช้ 15 ข้อต่อ [21], กราฟความถี่ของ Bag of Skeleton Codewords [22], ข้อมูลเส้นทางเคลื่อนที่ของโครงร่างแบบกระดูก 3 มิติ (3D Skeleton) [23], ฟีเจอร์โพส (Posture) จากโครงร่างแบบกระดูกและข้อต่อ 3 มิติ [24], โครงร่างแบบกระดูกโดยใช้ HMMs เพื่อสังเกตส่วนที่ถูกบดบังและสูญหายไป [25] ผลลัพธ์ของฟีเจอร์นี้จะมีแบบจำลองที่ค่อนข้างชัดเจนตามลักษณะของมนุษย์ แม้ว่าจะต้องแลกมาด้วยการคำนวณที่ซับซ้อน

ในการสร้างข้อต่อหรือโครงร่างกระดูก ซึ่งให้ผลความถูกต้องที่ค่อนข้างดี โดยต้องใช้การติดตามและทำนาย



ภาพประกอบที่ 1-2 ตัวอย่างฟีเจอร์ในลักษณะของข้อต่อและโครงกระดูก [18] [20]

(ข) การแสดงฟีเจอร์เป็นลักษณะการเคลื่อนไหวหรือการไหล (Motion/flow-based) เป็นลักษณะของฟีเจอร์โดยรวม โดยใช้ลักษณะการเคลื่อนไหว (Motion) หรือการไหล (Flow) ดังตัวอย่างฟีเจอร์ประเภทนี้ ได้แก่ ฟีเจอร์ของประวัติของปริมาตรการเคลื่อนไหว (Motion History Volume) ที่ไม่เปลี่ยนแปลง [26], Local Descriptor จากเส้นทางการเคลื่อนไหวของ Optical-flow [27], เส้นทางการเคลื่อนไหวย่อย ๆของ KLT Motion-based [28], Divergence-Curl-Shear Descriptor [29], ฟีเจอร์แบบผสมของ Contour และ Optical-flow [30], ประวัติของการเคลื่อนไหว (History of Motion) และ Optical-flow [31], ฟีเจอร์ชุดของการเคลื่อนไหว (Motion) หลายระดับ [32], ฟีเจอร์โปรเจคชั่นพลังงานสะสมของการเคลื่อนไหว (Accumulated Motion Energy) [33], พีระมิดของข้อมูลเชิงพื้นที่และเวลาจาก Motion Descriptor [34], ฟีเจอร์ Motion และ Optical-flow-based ด้วยกันกับ Markov Random Field สำหรับการทำนายส่วนที่ถูกบดบัง [35] วิธีการเหล่านี้จะไม่ต้องการการลบพื้นหลังที่ถูกต้องแม่นยำมาก แต่จะมาพร้อมกับข้อมูลที่ไม่คงที่ ซึ่งต้องใช้วิธีการหรือ Descriptor เพื่อที่จะเข้ามาจัดการ



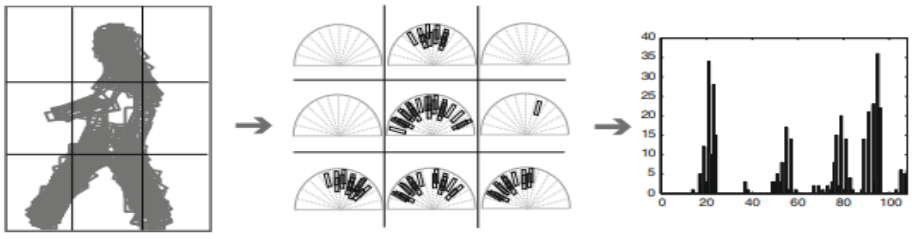
ภาพประกอบที่ 1-3 ตัวอย่างฟีเจอร์ Motion History Volume ในลักษณะการเคลื่อนไหว [26]

(ค) การแสดงฟีเจอร์ที่เป็นปริมาตรของวัตถุที่สัมพันธ์กับพื้นที่และเวลา (Space-time Volume-based) เป็นการสร้างแบบจำลองที่เป็นชุดของเงา รูปร่าง หรือพื้นผิว โดยจะใช้ช่วงเวลาหนึ่งของเฟรม เพื่อที่จะสร้างแบบจำลอง เช่น ฟีเจอร์เชิงพื้นที่และเวลาของ Silhouettes จาก Shape History Volume [36], ฟีเจอร์เชิงพื้นที่และเวลาของคุณสมบัติทางเรขาคณิตจากปริมาตรที่ต่อเนื่องกัน [37], ฟีเจอร์เชิงพื้นที่และเวลาของรูปร่างจาก 3D Point Cloud [38], ฟีเจอร์ชุดของรูปทรงจากพื้นที่และเวลาของ 3D Binary Cube [39], ฟีเจอร์พื้นที่และเวลาที่ไม่เปลี่ยนไปตามการหมุนในสามมิติ (Affine-invariant) [40], ฟีเจอร์เชิงพื้นที่และเวลาของ Micro Volume โดยใช้ Binary Silhouette [41], พื้นที่เชิงรวม (Integral Volume) ของขอบลำตัวมองเห็น (Visual-hull) และประวัติของปริมาตรการเคลื่อนไหว (Motion History Volume) [42], ปริมาตรที่มีลักษณะเด่น (Saliency Volume) จากองค์ประกอบของความสว่าง, สี และมุม [43] โดยวิธีการเหล่านี้จะได้มาซึ่งข้อมูลที่มีรายละเอียดที่สูง แต่ต้องแลกมากับข้อมูลฟีเจอร์ที่มีหลายมิติและขนาดใหญ่ ซึ่งต้องอาศัยความแม่นยำในการแยกภาพของคนออกจากพื้นหลัง เพื่อให้ได้มาซึ่งข้อมูลที่น่าเชื่อถือ

Space-time Shape [36] Micro-volume [41]

ภาพประกอบที่ 1-4 ตัวอย่างฟีเจอร์ลักษณะปริมาตรของวัตถุที่สัมพันธ์กับพื้นที่และเวลา

(ง) การแสดงฟีเจอร์ในรูปแบบของกริด (Grid-based) จะเป็นการแบ่งบริเวณที่เป็นตัวบุคคลที่ผู้วิจัยสนใจลงในเซลล์ (Cell), กริด (Grid) หรือบลอค (Block) ซึ่งจะบีบอัดความจำเพาะของข้อมูล รวมรวมข้อมูล การแทนข้อมูลแบบใหม่ หรือมีการดึงลักษณะเด่นไว้ในแต่ละพื้นที่ (Encode) อย่างเช่น ฟีเจอร์กริดของกราฟสะสมความถี่เชิงรูปร่างสี่เหลี่ยม (Histogram Oriented of Rectangular) [44], ตัวอธิบายคุณลักษณะของการไหล (Flow Descriptor) จากเซลล์ขนาดเล็กหลายเซลล์ที่สัมพันธ์กับพื้นที่และเวลา (Spatial-temporal Small Cells) [45], กราฟสะสมความถี่ของ Local Binary Pattern ในพื้นที่กริด [46], กริดรูปสี่เหลี่ยมของ Optical-flow [47], การสร้างฟีเจอร์ Codewords สำหรับกราฟสะสมความถี่เชิง Gradients (Histogram Oriented of Gradient) และ Optical-flow [48], จุด 3D ที่น่าสนใจ (Interest Point) ภายในหน้าต่างหลายขนาด [49], กราฟสะสมความถี่ของการเคลื่อนไหวของ Gradients [50], การรวมกันของประวัติการเคลื่อนไหว(Motion History), Local Binary Pattern และสำหรับกราฟสะสมความถี่เชิง Gradients [51] โดยวิธีการเหล่านี้จะมีความซับซ้อนในการสร้างแบบจำลองไม่มาก แต่ต้องแลกมากับการได้ข้อมูลที่ซ้ำกัน หรือข้อมูลที่ไม่มีนัยสำคัญในบางตำแหน่ง ซึ่งในงานวิจัยฉบับนี้ได้เลือกใช้ฟีเจอร์แบบกริดที่เป็นเลเยอร์ในแนวตั้ง เพื่อใช้ในการสร้างแบบจำลองมนุษย์ที่ใช้ในการรู้จำท่าทางโดยการฟิวชันฟีเจอร์ระดับล่างหลายมุมมอง ซึ่งทำให้ได้แบบจำลองที่มีความซับซ้อนน้อย มีประสิทธิภาพในการรู้จำได้ดีในหลาย ๆมุมมองที่แตกต่างกัน



ภาพประกอบที่ 1-5 ตัวอย่างฟีเจอร์รูปแบบของกริด [44]

นอกจากนี้ ในปัจจุบันได้มีการนำการเรียนรู้เชิงลึก (Deep Learning) มาใช้ในการเรียนรู้จดจำภาพต่าง ๆ ซึ่งได้กลายมาเป็นวิธีการพื้นฐานที่ใช้กันอย่างแพร่หลายเพื่อการรู้จำสิ่งที่มองเห็นได้ (Visual-based Recognition) เนื่องจากว่าการเรียนรู้เชิงลึกจะสามารถเรียนรู้ฟีเจอร์ที่ได้มาจากข้อมูลดิบ โดยใช้การสร้างลำดับชั้นต่าง ๆ (Hierarchically Layers) ซึ่งสามารถแทนข้อมูลของ

ฟีเจอร์ด้วยระดับต่าง ๆที่มีความซับซ้อนได้อย่างอัตโนมัติ ซึ่งก็มีงานวิจัยที่เกี่ยวข้องกับการรู้จำท่าทางจากการที่ได้ทบทวน ซึ่งมีตัวอย่างงานวิจัยที่ใช้การเรียนรู้เชิงลึกเพื่อรู้จำท่าทางของมนุษย์ดังนี้ การประยุกต์ใช้งาน Convolutional Neural Network กับข้อมูลภาพสามมิติแบบเต็มตัว และฟีเจอร์ดิกชินนารี่ของข้อมูลท่าทาง (Pose Dictionary Features) [52], การเรียนรู้ท่าทางของผู้ขับขี่จากภาพหลายมุมมองโดยใช้ Convolutional Neural Network ร่วมกับ Long Short Term memory [53], การรู้จำท่าทางหลายมุมมองด้วย Autoencoder สำหรับฟีเจอร์ที่ไม่เปลี่ยนไปตามมุมมอง (View-invariant Feature) [54] เป็นต้น

อย่างไรก็ตามงานวิจัยของผู้วิจัยได้เน้นไปที่การฟิวชันข้อมูลในระดับฟีเจอร์จากหลายมุมมองโดยใช้แบบจำลองที่สร้างขึ้นโดยเฉพาะ จึงยังจำเป็นต้องใช้วิธีการเรียนรู้แบบเดิมเพื่อรู้จำท่าทาง ซึ่งมีความเร็วในการประมวลผลที่รวดเร็วมากตามขอบเขตของงานวิจัยที่เน้นการประมวลผลแบบทันเวลา

1.3.2 การรู้จำท่าทางจากการฟิวชันข้อมูลระดับสูงจากหลายมุมมอง (Multi-view High-Level Fusion for Action Recognition)

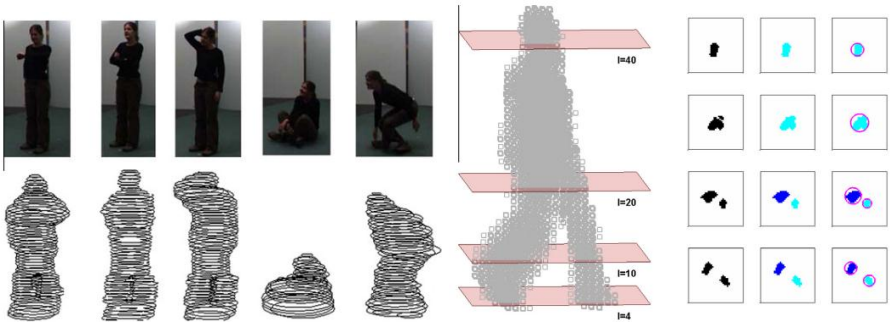
ในการฟิวชันข้อมูลในระดับคำตอบจากหลายมุมมองนั้น การรู้จำท่าทางจะเป็นการนำข้อมูลคำตอบของท่าทางมาเพื่อตัดสินใจ ชั่งน้ำหนัก วัดค่าความน่าเชื่อถือของคำตอบจากการทำนายคำตอบในมุมมองเดี่ยวต่าง ๆ ซึ่งงานวิจัยของ M.A. Naiel และคณะ [55] ได้นำเสนอการฟิวชันข้อมูลในระดับคำตอบโดยใช้ PCA 2 มิติ (2DPCA) และnearest neighbor เพื่อรู้จำท่าทางในแต่ละมุมมอง จากนั้นใช้เทคนิคการโหวตโดยเสียงส่วนใหญ่ (Majority Voting Technique) โดยอนุมานว่าท่าทางแต่ละมุมมองมีความสอดคล้องกันโดยที่ไม่สามารถมองข้ามได้ หากคะแนนโหวตยังไม่เป็นที่พอใจ ระบบจะเลือกการตัดสินใจของมุมมองที่มีค่าความผิดพลาดของการทำนายน้อยที่สุด และงานวิจัยของ A. Iosifidis และคณะ [56] ได้ใช้การโหวตท่าทางจากเสียงส่วนใหญ่ (Majority Voting Technique) จากหลายมุมมอง ร่วมกับ Bayesian Framework เพื่อรู้จำท่าทางจากหลายมุมมอง โดยใช้ฟีเจอร์แผนผังต้นแบบของท่า (Posture Prototype Map) ซึ่งจัดข้อมูลด้วย Self-organizing Map เพื่อรู้จำท่าทางในมุมมองเดี่ยวร่วมกับ Traditional Neural Network (TNN) ก่อนนำคำตอบจากการทำนายคำตอบในมุมมองเดี่ยวต่าง ๆมาฟิวชันข้อมูลระดับสูงจากหลายมุมมอง

1.3.3 การรู้จำท่าทางจากการฟิวชันข้อมูลระดับฟีเจอร์จากหลายมุมมอง (Multi-view Feature-Level Fusion for Action Recognition)

แม้ว่าการรู้จำท่าทางจากฟีเจอร์แบบต่าง ๆจากมุมมองเดี่ยวจะให้ผลที่ดีพอสมควร แต่ในการรู้จำท่าทางในการใช้งานจริงนั้น จะต้องมีความทนทานต่อการเปลี่ยนแปลงของข้อมูลในเรื่อง

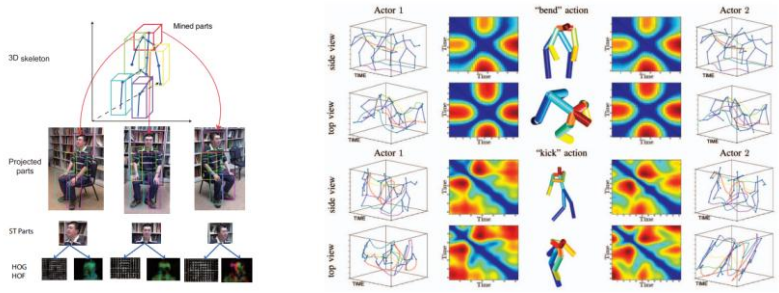
ของมุมมองด้านต่าง ๆที่กล้องจับภาพของมนุษย์ และมุมมองการติดตั้งกล้อง รวมไปถึงการจัดการกับปัญหา เช่น ซ้อนทับกันของบุคคลหรือการถูกบดบัง, การขาดข้อมูลภาพบางส่วนจากสิ่งรบกวน และการได้บริเวณที่สังเกตการณ์ได้มากขึ้น ซึ่งการรู้จำท่าทางจากหลายมุมมองจะสามารถแบ่งประเภทการรู้จำท่าทางจากการสร้างแบบจำลองจากการฟิวชันข้อมูลในมุมมองต่าง ๆของฟีเจอร์ในลักษณะของ 2 มิติ และ3 มิติ

สำหรับตัวอย่างงานวิจัยที่ใช้ 2 มิติ ได้แก่ การแทนข้อมูลด้วยแบบจำลองมนุษย์แบบเลเยอร์เชิงกลม [57], Bag of Visual-words โดยใช้จุดที่น่าสนใจเชิงพื้นที่และเวลาสำหรับแบบจำลอง [58], การแทนข้อมูลด้วย Masks และ Movements ของท่าทางที่ไม่แปรเปลี่ยนไปตามมุมมอง [59], ฟีเจอร์ R-transform [60], พื้นที่ของฟีเจอร์โครงร่างมนุษย์ด้วย PCA [61], คุณลักษณะเฉพาะของฟีเจอร์ระดับต่ำของมนุษย์ [62], การรวมกันของกราฟสะสมความถี่ Optical Flow และ Bag of Interest Point Word [63], Contour-based และ Uniform Local Binary Pattern [64], การใช้ฟีเจอร์หลายรูปแบบด้วย Keyposes Learning [65] ฟีเจอร์โครงของ Contours ที่ถูกลดมิติของข้อมูล [66], ฟีเจอร์แผนผังของท่าทางโดยใช้การวิเคราะห์จาก LDA บนภาพของท่าทางจากหลายมุมมอง [67]



ภาพประกอบที่ 1-6 ตัวอย่างงานวิจัยแบบ 2D ที่สร้างแบบจำลองของมนุษย์โดยใช้เลเยอร์ที่แสดงโดยวงแหวนที่ได้จากการประมาณค่า Voxel [57]

สำหรับตัวอย่างงานวิจัยที่ใช้ 3 มิติ โดยที่แบบจำลองตัวบุคคลถูกสร้างขึ้นใหม่ (Reconstruction) หรือจำลองขึ้นเป็นฟีเจอร์ใหม่ ซึ่งเป็น 3 มิติ จากหลายมุมมอง อย่างเช่น ฟีเจอร์พีระมิดของ Bag of Spatial-Temporal Descriptor และ Part-based ด้วยการเรียนรู้แบบ Induced Multi-task [68], ฟีเจอร์กราฟตรรกะเชิงพื้นที่และเวลาด้วยชิ้นส่วนต่าง ๆที่ถูกอธิบายไว้แล้ว [69], ฟีเจอร์ความคล้ายคลึงกันของรูปร่างเชิงเวลาในวีดีโอสามมิติ [70], ฟีเจอร์ Circular FFT จาก Convex Shape [71], ฟีเจอร์ Bag of Multiple Temporal Self-similarities [72], ฟีเจอร์ Circular Shift Invariance of DFT จากการเคลื่อนไหว [73] ซึ่งวิธีการที่ได้กล่าวมาข้างต้นนั้น จะสร้างแบบจำลองข้อมูลเชิงพื้นที่และเวลา (Temporal-spatial) ที่สามารถจะเพิ่มความแม่นยำ ซึ่งจะเพิ่มความถูกต้องในการรู้จำ แต่จะมีความซับซ้อนที่เพิ่มมากยิ่งขึ้น



3D Geometric parts [69] Multi-view 3D MoCap & SSMs [72]

ภาพประกอบที่ 1-7 ตัวอย่างงานวิจัยที่สร้างแบบจำลองของมนุษย์แบบ 3D [69] [72]

1.3.4 การเปรียบเทียบกับงานวิจัยอื่นที่เป็นการรู้จำท่าทาง (Comparison of Action Recognition Research)

จากตารางที่ 1-1 จะทำการเปรียบเทียบข้อเด่นและข้อด้อยของแต่ละงานวิจัยที่เกี่ยวเนื่องจากการรู้จำท่าทาง โดยมีบรรทัดฐานเกี่ยวข้องกับการนำไปใช้งานจริง โดยในการใช้งานเซ็นเซอร์ความเฉื่อยจะต้องมีการติดอุปกรณ์อยู่บนร่างกายเพื่อที่จะสามารถตรวจจับท่าทางโดยทำได้ทุกที่ที่มีอุปกรณ์ติดอยู่ แต่จะมีปัญหาเรื่องความไม่สะดวกในการพกพาหรือติดอุปกรณ์เหล่านั้น โดยวิธีการนี้ให้ความเป็นส่วนตัวที่สูง แต่ก็ยังมีความซับซ้อนอยู่บ้าง ส่วนในการใช้การมองเห็นจากกล้องสี ไม่จำเป็นต้องมีอุปกรณ์ใด ๆ มาติดอยู่กับตัว ซึ่งทำให้เกิดความรำคาญ ถึงแม้ว่าจะมีความซับซ้อนน้อยในการเก็บข้อมูลจากกล้อง แต่ก็ยังขาดความเป็นส่วนตัวจากข้อมูลภาพสีที่นำมาวิเคราะห์ ถัดมาคืองานที่ใช้ภาพความลึกจะให้ความเป็นส่วนตัวที่สูงกว่า แต่ก็ยังมีปัญหาบางอย่างซึ่งต้องใช้กล้องจากหลายมุมมองเพื่อเข้ามาจัดการบางข้อจำกัด เช่น การมองเห็นมุมมองที่กว้างกว่า, หลีกเลี่ยงปัญหาการขาดข้อมูลจากสิ่งรบกวน

จากงานของผู้วิจัยจะเห็นได้ว่าจะค่อนข้างมีประสิทธิภาพ เนื่องจากว่าแบบจำลองมีความซับซ้อนที่น้อย เน้นความเป็นส่วนตัวในการสังเกตการณ์ที่สูง เนื่องจากใช้ภาพความลึกและคุณสมบัติอื่น ๆ เช่น ความยืดหยุ่นสามารถเพิ่มขยายระบบได้ และความทนทานซึ่งมีระดับที่ค่อนข้างดี หรือไม่น้อยกว่า และไม่ต้องการการสอบเทียบค่าของกล้องในการติดตั้ง (Calibration)

ซึ่งแม้ว่างานของผู้วิจัยจะใช้กล้องสองตัวเพื่อที่จะได้ข้อมูลหลายมุมมอง แต่ก็ต้องมีการติดตั้งที่มีค่าเฉพาะบางอย่าง เช่น มุมที่กระทำกันระหว่างกล้องต้องมากกว่า 30°, ต้องติดตั้งกล้องสูงเหนือพื้นประมาณ 2 เมตร

ตารางที่ 1-1 เปรียบเทียบงานวิจัยที่รู้จำท่าทางบนพื้นฐานของการนำไปใช้งานจริง

ประเภทของงานวิจัย	ความยืดหยุ่นในการ calibrate	ความสามารถในการขยายจำนวนของมุมมอง	ความทนทานต่อมุมมองที่เปลี่ยนแปลงไป	เงื่อนไขในการใช้งานจริง		
				ความซับซ้อน	ความรำคาญจากการติดอุปกรณ์	ความเป็นส่วนตัว
Inertial sensor-based [5-16]	N/A	N/A	สูง	ขึ้นอยู่กับงาน	มี	สูง
RGB vision-based [17] [21] [25-51] [74-76][78]	น้อย-ปานกลาง	น้อย-ปานกลาง	น้อย-ปานกลาง	น้อย-ปานกลาง	ไม่มี	ไม่มี
Depth-based Single-View [4] [18-20] [22-24] [78]	น้อย-ปานกลาง	น้อย-ปานกลาง	น้อย-ปานกลาง	ขึ้นอยู่กับงาน	ไม่มี	ปานกลาง-สูง
RGB/Depth-based Multi-View [57-73]	น้อย-ปานกลาง	น้อย-ปานกลาง	น้อย-ปานกลาง	ขึ้นอยู่กับงาน	ไม่มี	ขึ้นอยู่กับงาน
งานวิจัยนี้ - Depth-based Multi-View	**ปานกลาง-สูง**	**ปานกลาง-สูง**	**สูง**	**ซับซ้อนน้อย**	**ไม่มี**	**สูง**

เมื่อเปรียบเทียบกับงานวิจัยอื่นที่เป็นการรู้จำท่าทางพื้นฐานของมนุษย์จากการใช้กล้องจาก ตารางที่ 1-2 แสดงให้เห็นถึงประสิทธิภาพโดยวิธีการของผู้วิจัย และเทคนิคการรู้จำท่าทางพื้นฐานของมนุษย์จากการมองเห็นอื่น ๆ โดยค่าเฉลี่ยความแม่นยำจากงานของผู้วิจัยมีค่าเท่ากับ 95.32% ซึ่งอยู่ระดับที่ใกล้เคียงกับงานอื่น ๆ โดยการรู้จำท่าทางท่าเดินและท่านั่งจะมีความถูกต้องมากที่สุด รองลงมาคือท่ายืนและท่านอนยังมีความถูกต้องที่ยังยอมรับได้อยู่ แต่ท่าที่ได้ค่าความถูกต้องน้อยที่สุดคือท่าก้ม แต่ก็ยังรู้จำได้ความถูกต้องสูงถึง 92.72% แม้ว่าจะไม่สามารถเปรียบเทียบด้านความทนทานและประสิทธิภาพด้านเวลาได้ เนื่องจากขาดข้อมูล และแต่ละ

งานวิจัยไม่ได้ใช้ชุดข้อมูลทดสอบชุดเดียวกัน งานวิจัยของผู้วิจัยจะทดสอบและมุ่งเน้นไปที่วิธีการรู้จำท่าทางที่มีความทนต่อมุมมองที่เปลี่ยนแปลงไป และยังประมวลผลวิเคราะห์ท่าทางได้ 63 เฟรมต่อวินาที

ตารางที่ 1-2 การเปรียบเทียบกับงานวิจัยอื่นที่เป็นการรู้จำท่าทางพื้นฐานของมนุษย์จากการมองเห็น

งานวิจัย	อัตราความแม่นยำ (%)					
	ยืน	เดิน	นั่ง	ก้ม	นอน	ค่าเฉลี่ย
P. Chawalitsittikul และคณะ [4]	98.00		93.00	94.10	98.00	95.78
N. Noorit และคณะ [74]	99.41	80.65	89.26	94.35	100.0	92.73
M. Ahmad และคณะ [75]	-	89.00	85.00	100.0	91.00	91.25
C. H. Chuang และคณะ [76]	-	92.40	97.60	95.40	-	95.80
G. I. Parisi และคณะ [77]	96.67	90.00	83.33	-	86.67	89.17
N. Sawant และคณะ [78]	91.85	96.14	85.03	-	-	91.01
งานของผู้วิจัย	99.31		98.59	92.72	90.65	95.32

*หมายเหตุ แต่ละงานวิจัยไม่ได้ใช้ชุดข้อมูลทดสอบชุดเดียวกัน

1.3.5 การติดตามตัวบุคคลจากหลายมุมมอง (Multi-view People Tracking)

การติดตามและจดจำตัวบุคคลเป็นสิ่งที่สำคัญอย่างหนึ่งในระบบเฝ้าระวังอัจฉริยะต่าง ๆ ที่ทำงานเกี่ยวกับการวิเคราะห์และติดตามพฤติกรรมของมนุษย์ ซึ่งเมื่อบุคคลเมื่อเข้ามาในบริเวณที่ระบบทำการวิเคราะห์ จะต้องสามารถติดตามและบอกตำแหน่ง หรืออาจจะมีการทำนายตำแหน่งไปด้วยก็ได้ ซึ่งการติดตามตัวบุคคลในบางงานวิจัยจะใช้ขั้นตอนวิธีที่รู้จักกันอย่างแพร่หลาย เช่น Kalman Filtering, Particle Filtering มาประยุกต์ใช้ในการติดตามและทำนายตำแหน่งของบุคคล ในขณะเดียวกันงานวิจัยอีกประเภทจะใช้ขั้นตอนวิธีพิเศษที่สร้างขึ้นมาเพื่อติดตามและทำนายตำแหน่งของบุคคลโดยเฉพาะ

ซึ่งในการติดตามและทำนายตำแหน่งของบุคคลที่ใช้ขั้นตอนวิธีพิเศษจากหลายมุมมองนั้น มักจะใช้การจับคู่คุณลักษณะของบุคคลที่สอดคล้องกันระหว่างมุมมอง รวมไปถึงการใช้พิกัดตำแหน่งของบุคคลจากข้อมูลในสามมิติหรือข้อมูลการวางตำแหน่งของกล้องต่าง ๆ เพื่อที่จะแยกแยะบุคคลต่าง ๆที่กำลังติดตามอยู่ โดยสามารถแบ่งประเภทของการติดตามเป็นกลุ่มใหญ่ได้ 2 ประเภท คือ (1) ประเภทที่ใช้การวัดค่าของกล้อง (Calibrated Camera) (2) ประเภทที่ไม่ใช้การวัดค่าของกล้องหรือกระทำโดยอัตโนมัติ (Uncalibrated Camera)

ในงานวิจัยประเภทที่ใช้การวัดค่าของกล้องมักจะใช้ข้อมูล 3D และการวางตำแหน่ง (3D Information and Alignment) เป็นข้อมูลในการติดตาม โดยจะต้องมีกระบวนการที่ต้องทราบการติดตั้งกล้องระหว่างมุมมองมาก่อน (Prior Calibration) หรือมีแบบจำลองสามมิติจากการให้ข้อมูล หรือมีขั้นตอนหรืออุปกรณ์ที่จะช่วย Calibrate ก่อน ซึ่งงานวิจัยประเภทนี้จะมีความทนทานต่อสภาพแวดล้อมในการทำงานค่อนข้างสูง แต่ต้องแลกมาด้วยความซับซ้อนในการประมวลผลและต้องทำการ Calibrate ระบบก่อนหน้า ตัวอย่างเช่น Fleuret และคณะ [79] ได้นำเสนอ Probabilistic Occupancy Map (POM) เพื่อประมาณตำแหน่งของบุคคลจากแบบจำลองสีและความเคลื่อนไหว โดยใช้ขั้นตอนวิธี Viterbi, Zhao และคณะ [80] ได้ใช้กล้องส-เตอริโอหลายตัวเพื่อตรวจจับบุคคลใน Ground Plane Projection ด้วยวิธีมองอย่างรวดเร็ว (Rapid Skimming) โดยบุคคลจะถูกติดตามในกล้องเดี่ยวจะใช้ข้อมูลรูปร่าง รูปลักษณ์ และความแตกต่างกันของพื้นที่ยืนอยู่ จากนั้นจะถูกจับคู่ไปยังมุมมองอื่น ๆโดยใช้วิธีการผสานกันของ ground-based ในข้อมูลเชิงพื้นที่และเวลา, Mittal และคณะ [81] ได้นำเสนอ M_2Tracker เพื่อติดตามบุคคลที่อยู่ชิด ๆกัน โดยแบบจำลองบุคคลโดยสี ความน่าจะเป็นจากความสูงและความกว้าง ซึ่งจะใช้ฉายความน่าจะเป็นของตัวบุคคลลงบนแผ่นพื้น (Ground Plane) ที่ได้จากการหาความสัมพันธ์กันของจุดในสามมิติที่ได้จากความสอดคล้องกันของเส้น Epipolar

งานวิจัยประเภทที่ไม่ใช้การวัดค่าของกล้องหรือกระทำโดยอัตโนมัติ (Uncalibrated Camera) มักจะใช้การจับคู่กันของฟีเจอร์ที่สอดคล้องกันระหว่างมุมมอง การจับคู่ตำแหน่งที่อยู่ของมุมมองจากกล้อง ตัวอย่างเช่น Muñoz-Salinas และคณะ [82] ได้ใช้ Evidential Particle Filter ที่เป็นส่วนขยายของ Bayesian Particle Filters กับ Dempster-Shafer Theory ซึ่งถูกปรับ

มาจาก Evidential Filter เพื่อการผสานข้อมูลในการติดตามตัวบุคคลจากหลายมุมมอง จากข้อมูลฉากหน้าที่เป็นตัวบุคคล (Foreground) สี และรูปร่าง, Hu และคณะ [83] ได้ใช้การจับคู่ระหว่างหลาย ๆมุมมองด้วยการหาความเป็นไปได้ที่สอดคล้องกันของแบบจำลองความเหมือนกันของฟีเจอร์แกนหลัก (Principle Axis), Chang และคณะ [84] ได้ใช้วิธีการรวมหลายฟีเจอร์ในการติดตาม ได้แก่ Epipolar Geometry, Homography, Landmark, ความสูง และสี เพื่อหาความสอดคล้องกันในหลายมุมมองโดยใช้ Bayesian Modality Fusion และ Naive Bayes, Khan และคณะ [85] ได้ใช้การค้นหาขอบของ Field of View เพื่อหาความสัมพันธ์ในมุมกล้องต่าง ๆของบุคคลเดียวกัน

1.4 ขอบเขตของการวิจัย

1.4.1 เจาะจงเฉพาะท่าทางพื้นฐานของมนุษย์ 5 ท่าทาง ได้แก่ การยืน / การเดิน การนั่ง การก้ม และการนอน

1.4.2 การทดลองจะทำในสภาพแวดล้อมภายในอาคารที่ถูกควบคุม โดยใช้มุมมองที่ซ้อนทับกันของกล้องที่อยู่นิ่งกับที่

1.4.3 ข้อมูลที่นำมาวิเคราะห์จะเป็นข้อมูลภาพสีและความลึกเท่านั้น

1.4.4 มุ่งเน้นไปที่การติดตามตัวบุคคลหลายบุคคลและการรู้จำท่าทาง

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 สามารถนำไปประยุกต์ใช้ในระบบรักษาความปลอดภัย อย่างเช่น ตรวจจับพฤติกรรมที่ผิดปกติ และเหตุการณ์ต่าง ๆ

1.5.2 สามารถนำไปประยุกต์ใช้ในระบบดูแลด้านสุขภาพ อย่างเช่น ระบบดูแลเฝ้าระวังผู้สูงอายุระยะไกล และงานด้านการส่งเสริมสุขภาพ

บทที่ 2
ทฤษฎีและหลักการ

2.1 กระบวนการประมวลก่อนในขั้นต้น (Pre-processing)

กระบวนการประมวลก่อนในขั้นต้น (Pre-processing) จะส่วนของการปรับปรุงภาพความลึกจากกล้อง เนื่องจากภาพความลึกที่ได้มาอาจจะมีหลุมที่ไม่มีค่าความลึก และมีค่าความแปรปรวนในบางบริเวณที่เป็นพื้นที่เล็ก ๆคล้ายกับสัญญาณรบกวนเกลือและพริกไทย จึงจำเป็นต้องใช้ตัวกรองสัญญาณแบบมัธยฐาน (Median Filter) เพื่อปรับปรุงภาพในขั้นต้นก่อนกระบวนการตรวจจับบุคคลจากการจับภาพเคลื่อนไหว ซึ่งหลังจากการจับภาพเคลื่อนไหวก็จะเกิดสัญญาณรบกวนแบบกลุ่มจุดที่เกิดจากความแปรปรวนที่มีค่าสูงของความลึกซึ่งกระบวนการจับภาพเคลื่อนไหวไม่สามารถแบบจำลองบริเวณนั้นเป็นพื้นหลังในบางเฟรมได้ จึงจำเป็นต้องใช้กระบวนการทางสัณฐานวิทยา (Morphological Processing) เพื่อลบสิ่งรบกวนเหล่านี้

2.1.1 ตัวกรองสัญญาณแบบมัธยฐาน (Median Filter)

ตัวกรองสัญญาณแบบมัธยฐาน [86] นั้นเป็นตัวกรองที่มีความสามารถทำให้สัญญาณภาพเบลอเพื่อลดสัญญาณรบกวน หรือกล่าวคือเป็น Smoothing Filter ซึ่งใช้กระบวนการทางสถิติเพื่อปรับค่าพิกเซลให้มีความราบเรียบนั่นเอง โดยที่มีพื้นฐานมาจากค่ามัธยฐาน ซึ่งเป็นการหาตัวเลขที่อยู่กึ่งกลางของกลุ่มตัวเลข นั่นคือ ตัวเลขจำนวนครึ่งหนึ่งจะมีค่ามากกว่าค่ามัธยฐาน และอีกครึ่งหนึ่งจะมีค่าน้อยกว่าค่ามัธยฐาน ตัวอย่างเช่น ค่ามัธยฐานของ 1, 3, 3, 5, 7, 12 คือ 4 (ในกรณีเป็นจำนวนคู่ หากเป็นจำนวนคี่ ค่ามัธยฐานจะเป็นเลขที่อยู่ตรงกลางพอดี)

โดยแนวคิดของตัวกรองสัญญาณแบบมัธยฐานจะมีความใกล้เคียงกับตัวกรองสัญญาณแบบค่าเฉลี่ย ซึ่งจะใช้วิธีการพิจาณาแต่ละพิกเซลรอบ ๆบริเวณที่กำลังประมวลผลเพื่อจะเปลี่ยนให้ตรงกลางและรอบ ๆมีค่าใกล้เคียงกัน แต่ตัวกรองสัญญาณแบบมัธยฐาน จะนำค่าในพิกเซลที่ตัวกรองกำลังพิจารณามาเรียงกันแล้วนำค่าที่อยู่กึ่งกลางที่แทน ณ พิกเซลตรงกลางตามการหาค่ามัธยฐาน ซึ่งตัวกรองจะเป็น 1D หรือ 2D ก็ได้ ซึ่งแสดงตัวอย่างวิธีการทำงานของตัวกรองดังภาพประกอบที่ 2-1 และ 2-2

126	128	125	129	130
124	123	149	124	145
118	111	118	122	134

Pixel Values (sorted) : 123 124 149

Median Value : 124 (Replace in Center)

ภาพประกอบที่ 2-1 ตัวอย่างการทำงานของตัวกรองสัญญาณแบบมัธยฐาน 1D

123	124	131	129	126
126	128	125	129	130
124	123	149	125	145
118	111	118	122	134
110	115	114	123	125

Pixel Values (sorted) : 111 118 122 123 125 125 128 129 149

Median Value : 125 (Replace in Center)

ภาพประกอบที่ 2-2 ตัวอย่างการทำงานของตัวกรองสัญญาณแบบมัธยฐาน 2D

2.1.2 กระบวนการทางสัณฐานวิทยาของภาพ (Morphological Image Processing)

กระบวนการทางสัณฐานวิทยา [87] จะใช้เป็นกระบวนการที่จะเปลี่ยนแปลงรูปร่างหรือโครงสร้างของวัตถุในภาพ โดยเฉพาะอย่างยิ่งกับวัตถุที่เป็นภาพขาวดำแบบไบนารี่ โดยมีข้อมูลภาพตั้งต้น และ Structuring Element หรือ Kernel ที่ มากระทำกันทางสัณฐานวิทยา เช่น การกัดกร่อนภาพ (Erosion), การพอกภาพ (Dilation), การเปิดภาพ (Opening), การอุดภาพ (Closing), การทำให้เหลือแต่โครง (Skeletonization) เป็นต้น เพื่อให้ได้รูปร่างใหม่ของวัตถุในภาพตามต้องการ ดังตัวอย่างในภาพประกอบที่ 2-3

(ก) (ข) (ค) (ง) (จ) (ฉ)

ภาพประกอบที่ 2-3 ตัวอย่างการทำงานของกระบวนการทางสัณฐานวิทยาของภาพ (ก)ภาพต้นฉบับ (ข)ภาพที่ผ่านการกัดกร่อนภาพ (ค)ภาพที่ผ่านการพอกภาพ (ง)ภาพที่ผ่านการเปิดภาพ (จ)ภาพที่ผ่านการอุดภาพ (ฉ)ภาพที่ผ่านการทำให้เหลือแต่โครง

ซึ่งโดยหลัก ๆแล้ว จะมีกระบวนการพื้นฐาน 2 กระบวนการ คือ การกัดกร่อนภาพ (Erosion), การพอกภาพ (Dilation) ซึ่งจะนำไปประกอบเป็นกระบวนการอื่น ๆได้ โดยการเปิดภาพ (Opening) จะประกอบไปด้วย การกัดกร่อนภาพ แล้วตามด้วย การพอกภาพตามลำดับ โดยผลลัพธ์จะทำให้ส่วนที่เชื่อมกันที่มีขนาดเล็กถูกเปิดออก สิ่งรบกวนขนาดเล็กที่เป็นจุดหายไป, ส่วนการอุดภาพ (Closing) จะประกอบไปด้วย การพอกภาพ แล้วตามด้วย การกัดกร่อนภาพตามลำดับ โดยผลลัพธ์จะทำให้ส่วนที่เป็นหลุมเป็นร่องที่มีขนาดเล็กถูกปิดลง

โดยในงานวิจัยฉบับนี้ได้เลือกใช้เฉพาะการกัดกร่อนภาพ (Erosion), การพอกภาพ (Dilation) ประกอบกันเพื่อปรับปรุงรูปร่างของวัตถุและลดสิ่งรบกวนในภาพ โดยทั้งหมดนี้มี

องค์ประกอบ 2 อย่าง คือ ภาพขาวดำแบบไบนารี่ และ Structuring Element หรือ Kernel ที่ มากระทำกันทางสัณฐานวิทยาแบบการแตะ (Hit) และการพอดี (Fit)

ซึ่ง Kernel จะมีเป็นไบนารี่เมทริกซ์ที่มีขนาดต่าง ๆกัน ตามการออกแบบ หากว่าต้องการกระทำกับวัตถุในภาพที่มีขนาดใหญ่ก็ต้องใช้เมทริกซ์ขนาดใหญ่ตามไปด้วยและเมทริกซ์จะมีค่าข้างในที่ต่างกันตามการออกแบบเพื่อให้ได้ผลลัพธ์ทางสัณฐานวิทยาของภาพตามต้องการ เช่น Rectangular Kernel, Cross-shaped Kernel, Elliptical Kernel เป็นต้น ตามตัวอย่างในภาพประกอบที่ 2-4

1	1
1	1

(ก)

0	1	0
1	1	1
0	1	0

(ข)

0	0	1	0	0
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
0	0	1	0	0

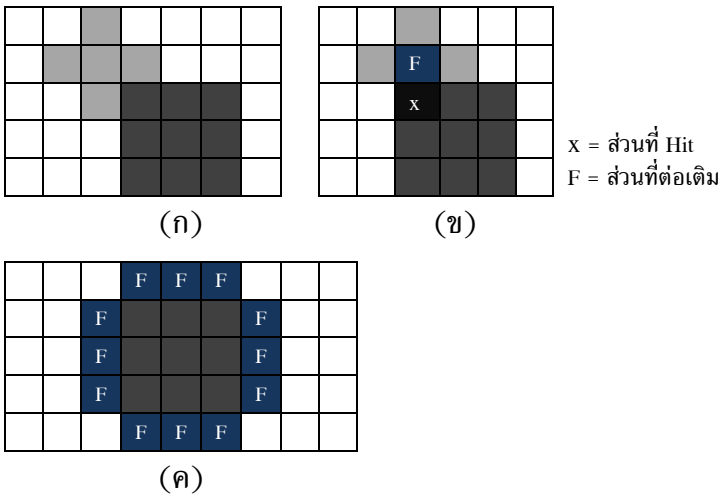
(ค)

ภาพประกอบที่ 2-4 ตัวอย่างของ Kernel ทางสัณฐานวิทยาของภาพ

(ก) Rectangular kernel 2x2 (ข) Cross-shaped Kernel 3x3 (ค) Elliptical kernel 5x5

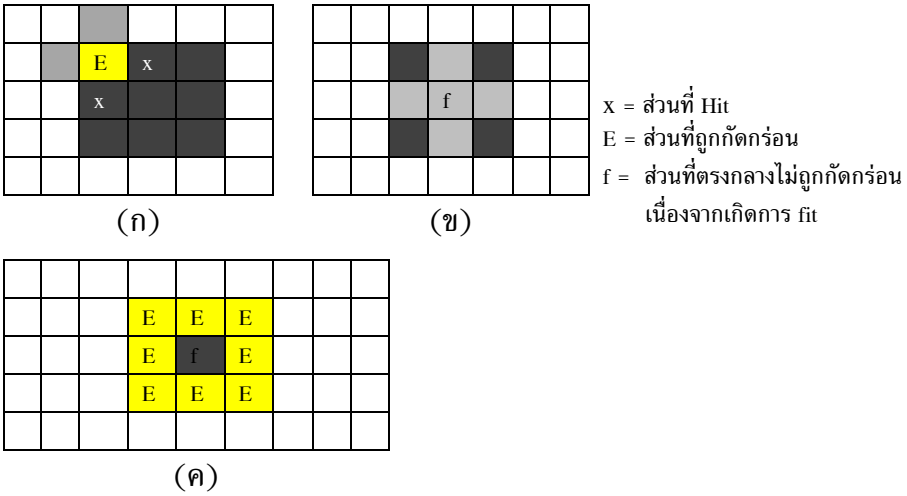
ซึ่งการดำเนินการกัดกร่อนภาพ (Erosion), การพอกภาพ (Dilation) จะเป็นการนำ Kernel ไปทับกับภาพตั้งแต่จุด Origin ซ้ายบนแล้วทำการพิจารณาการแตะ (Hit) หรือการพอดี (Fit) ในจุดนั้น ๆให้เสร็จแล้วจึงขยับ Kernel ไปทางขวาทีละ 1 พิกเซลจนสุดทางด้านขวาของภาพ แล้วขึ้นบรรทัดใหม่โดยเลื่อนมา 1 พิกเซล แล้วทำซ้ำแบบเดิมนี้ไปจนทั่วทั้งภาพ

โดยขั้นตอนของการพอกภาพ (Dilation) จะใช้การแตะ (Hit) เพื่อเป็นการต่อเติมภาพ ซึ่งการแตะ คือการที่การนำ Kernel ไปซ้อนกับภาพไบนารี่ ณ จุดใดจุดหนึ่ง แล้วปรากฏว่ามีค่าเลข 1 ของภาพไบนารี่ และ Kernel มาซ้อนทับกันเพียงจุดใดจุดหนึ่ง ซึ่งเมื่อเกิดการแตะสำหรับการพอกภาพก็จะพิจารณาให้มีการต่อเติมภาพ(เติม 1) ณ จุดตรงกลางของ Kernel ที่เกิดการแตะ ซึ่งหากไม่ปรากฏว่ามีค่าเลข 1 ของภาพไบนารี่ และ Kernel มาซ้อนทับกันเลยซักจุดจะเรียกว่าการพลาด (Miss) ซึ่งก็จะไม่เกิดการต่อเติมภาพ ดังตัวอย่างในภาพประกอบที่ 2-5



ภาพประกอบที่ 2-5 ตัวอย่างของการพอกภาพ (Dilation)
(ก) เกิดการพลาด Miss (ข) การแตะ Hit (ค) ผลลัพธ์สุดท้ายของการพอกภาพ

ส่วนการกัดกร่อนภาพ (Erosion) จะใช้การแตะ (Hit) เพื่อเป็นการกัดภาพ ซึ่งการแตะคือการที่การนำ Kernel ไปซ้อนกับภาพไบนารี่ ณ จุดใดจุดหนึ่ง แล้วปรากฏว่ามีค่าเลข 1 ของภาพไบนารี่ และ Kernel มาซ้อนทับกันเพียงจุดใดจุดหนึ่ง ซึ่งเมื่อเกิดการแตะสำหรับการกัดกร่อนภาพก็จะพิจารณาให้มีการกัดกร่อนภาพ(เติม0) ณ จุดตรงกลางของ Kernel ที่เกิดการแตะ ซึ่งหากว่ามีการซ้อนทับของเลขค่าเลข 1 ของ Kernel และภาพไบนารี่ในทุกจุด จะเรียกว่าการพอดี (Fit) ซึ่งก็จะไม่เกิดการกัดกร่อนภาพ เช่นเดียวกับการพลาด (Miss) ที่จะไม่มีอะไรเกิดขึ้นเลย ตัวอย่างการกัดกร่อนภาพจะแสดงในภาพประกอบที่ 2-6



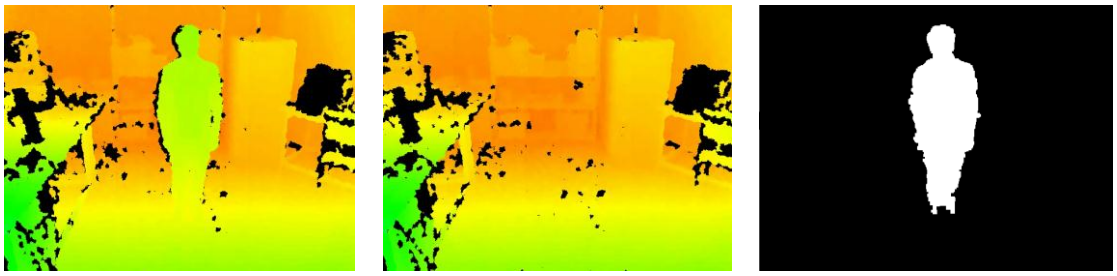
ภาพประกอบที่ 2-6 ตัวอย่างของการกัดกร่อนภาพ (Erosion)
(ก) เกิดการแตะ Hit (ข) เกิดการพอดี Fit (ค) ผลลัพธ์สุดท้ายของการกัดกร่อนภาพ

2.2 การตรวจจับความเคลื่อนไหว (Motion Detection)

การตรวจจับความเคลื่อนไหวในงานวิจัยนี้จะใช้การลบภาพออกจากแบบจำลองพื้นหลังที่ผสมผสานหลายเกาส์เซียน (Background Subtraction using Mixture of Gaussians Background Model) ซึ่งเป็นขั้นตอนวิธีที่จะแยกพื้นหลังที่ถูกแบบจำลองด้วยเกาส์เซียนหลายตัวกับฉากหน้าที่เป็นวัตถุที่เคลื่อนไหวในภาพ (Foreground) ออกจากกันร่วม โดยที่วิธีการนี้จะแบบจำลองแต่ละพิกเซลที่เป็นพื้นหลังโดยการผสมกันของกราฟการกระจายตัวของเกาส์เซียนเท่า K ตัว โดยที่ค่าน้ำหนักของการผสมกันจะแทนด้วยสัดส่วนเวลาที่สีนั้น ๆยังคงอยู่ในฉาก ซึ่งความน่าจะเป็นของค่าสีพื้นหลังคือสิ่งเดียวที่จะทำให้อยู่ได้นานขึ้นและคงที่มากขึ้น

โดยแนวความคิดเริ่มต้นของการตรวจจับความเคลื่อนไหวแบบการลบภาพออกจากแบบจำลองพื้นหลัง โดยจะมีการจำแบบจำลองของภาพพื้นหลังเอาไว้ (Background) เมื่อมีเฟรมที่ต้องการตรวจจับความเคลื่อนไหว (Motion Frame) ก็จะนำมาลบกับ ภาพพื้นหลัง ซึ่งก็จะได้ฉากหน้าที่เป็นส่วนที่มีความเคลื่อนไหว ดังที่แสดงตัวอย่างในภาพประกอบที่ 2-7

โดยมีหลายงานวิจัยได้นำเสนอวิธีการของแบบจำลองภาพพื้นหลัง ที่สามารถแก้ปัญหาต่าง ๆได้ เช่น ภาพพื้นหลังที่มีความซับซ้อน ภาพพื้นหลังที่มีการเปลี่ยนแปลงได้หลายรูปแบบ การเปลี่ยนแปลงของแสงเมื่อวัตถุเข้ามาในเฟรม เงาของวัตถุที่เคลื่อนไหวในภาพ ทำให้ค่าของจุดสีที่ในแบบจำลองภาพพื้นหลังเปลี่ยนแปลงไป



ภาพประกอบที่ 2-7 ตัวอย่างของการจับความเคลื่อนไหวแบบการลบภาพออกจากแบบจำลองพื้นหลัง

โดยงานวิจัยของ Graimson และ Stauffer [88] [89] โดยผู้เขียนได้นำเสนอการสร้างแบบจำลองภาพพื้นหลังในแต่ละพิกเซลโดยการมีหลาย K ของกราฟการกระจายตัวแบบเกาส์เซียนโดยที่ K มักจะมีจำนวนที่ 3-5 ตัว เพื่อจัดการกับพิกเซลของภาพพื้นหลังที่สามารถเปลี่ยนแปลงได้หลายค่า โดยเกาส์เซียนที่ต่างกันจะถูกสมมุติให้แทนค่าสีที่ต่างกัน ส่วนพารามิเตอร์ถ่วงน้ำหนักของการผสมจะแทนสัดส่วนของเวลาที่สีเหล่านั้นอยู่ในฉาก

โดยแต่ละพิกเซลในฉากจะถูกจำลองด้วยการผสมกันของหลาย K ของกราฟการกระจายตัวแบบเกาส์เซียน โดยความน่าจะเป็นที่พิกเซลหนึ่งนั้นจะประกอบด้วยค่าของ X_N ที่เวลา N ตามสมการที่ (2.1)

$$p(x_N) = \sum_{j=1}^{K} w_j \eta(x_N; \theta_j) \tag{2.1}$$

โดยที่ w_k คือ พารามิเตอร์ถ่วงน้ำหนักขององค์ประกอบเกาส์เซียนที่ k และ $\eta(x_N; \theta_j)$ คือการะกระจายตัวแบบปกติขององค์ประกอบเกาส์เซียนที่ k ซึ่งสามารถแทนได้ตามสมการที่ (2.2)

$$\eta(x_N; \theta_j) = \eta(x; \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{D}{2}} \left|\Sigma_k\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathrm{x}-\mu_k)^T \Sigma_k^{-1} (\mathrm{x}-\mu_k)} \tag{2.2}$$

โดยที่ μ_k คือ ค่ากลาง และ $\Sigma_k = \sigma_k^2 I$ คือ ค่าความแปรปรวนของเกาส์เซียนที่ k

โดยการกระจายตัว ณ กราฟ K จะถูกเรียงโดยมีพื้นฐานมาจากค่าความพอดีที่ w_k / σ_k และการกระจายตัวลำดับแรกที่ B จะถูกใช้เป็นแบบจำลองของพื้นหลังในฉากโดยที่ B จะหาค่าได้ดังสมการที่ (2.3)

$$B = \arg\min\left(\sum_{j=1}^{b} w_j > T) \right) \tag{2.3}$$

โดยที่ค่าขีดแบ่ง T คือ ค่าสัดส่วนน้อยที่สุดของแบบจำลองพื้นหลังซึ่งเป็นความน่าจะเป็นที่น้อยที่สุด โดยการลบพื้นหลังจะถูกทำ ที่พิกเซลที่มีค่าเบี่ยงเบนเกิน 2.5 จากที่กระจายตัวลำดับแรกที่ B และการกระจายตัวของเกาส์เซียนตัวแรกที่ตรงกันกับค่าที่ทดสอบจะถูกอัพเดทค่าโดยสมการที่ (2.4)

$$\begin{aligned}
&\hat{w}_k^{N+1} = (1-\alpha)\hat{w}_k^{N} + \alpha \hat{p}(\omega_k \mid \mathrm{x}_{N+1}) \\
&\hat{\mu}_k^{N+1} = (1-\alpha)\hat{\mu}_k^{N} + \rho(\mathrm{x}_{N+1}) \\
&\hat{\Sigma}_k^{N+1} = (1-\alpha)\hat{\Sigma}_k^{N} + \rho(\mathrm{x}_{N+1} - \hat{\mu}_k^{N+1})(\mathrm{x}_{N+1} - \hat{\mu}_k^{N+1})^T \\
&\rho = \alpha \eta(\mathrm{x}_{N+1}; \hat{\mu}_k^{N}, \hat{\Sigma}_k) \\
&\hat{p}(\omega_k \mid \mathrm{x}_{N+1}) = \begin{Bmatrix} 1 & : if\ \omega_k \equiv 1^{st}\ match\ gaussian \\ 0 & : otherwise \end{Bmatrix}
\end{aligned} \tag{2.4}$$

ซึ่ง P. KadewTraKulPong และคณะ [90] ก็ได้นำเสนอการประยุกต์การตรวจจับเงาเพิ่มเติมจากงานวิจัยที่ได้อธิบายไว้ข้างต้น โดยใช้การแทนข้อมูลแบบพื้นที่ของแม่สี (Chromatic Color Space) โดยที่ต้องพิจารณาแบบจำลองของสีที่สามารถแยกระหว่างแม่สีและองค์ประกอบของความสว่างได้ โดยใช้การเปรียบเทียบระหว่างพิกเซลที่ไม่ใช่พื้นหลังกับองค์ประกอบของพื้นหลังปัจจุบัน ถ้าความแตกต่างทั้งแม่สีและความสว่างอยู่ในค่าขีดเกณฑ์ จะถูกพิจารณาว่าเป็นเงา

ซึ่งใช้แบบจำลองสีที่ถูกนำเสนอโดย T. Horprasert และคณะ [91] โดยจะพิจารณาเวกเตอร์ตำแหน่งที่ค่ากลางของ RGB ที่พิกเซลพื้นหลังที่ (E), เส้นของแม่ที่คาดการณ์ไว้ ($\|E\|$), การบิดเบี้ยวของแม่สี (d) และค่าขีดเกณฑ์ความสว่าง (τ) โดยให้ค่าของพิกเซลที่ถูกพิจารณาเป็น I, ค่าการบิดเบี้ยวของความสว่าง เป็น a และค่าการบิดเบี้ยวของสี เป็น c จากแบบจำลองของพื้นหลังที่สามารถจะคำนวณได้ตามสมการที่ (2.5) และ (2.6)

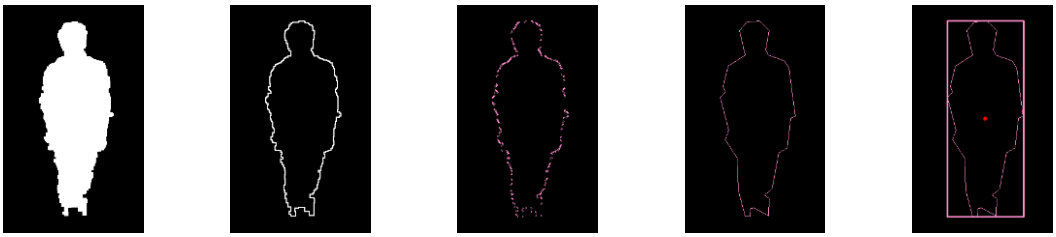
$$a = \arg\min(\mathrm{I} - \mathrm{zE})^2 \qquad (2.5)$$

$$c = \|I - aE\| \qquad (2.6)$$

และด้วยสมมุติฐานของการกระจายตัวทรงกลมแบบเกาส์เซียนในแต่ละองค์ประกอบที่รวมกัน ค่าเบี่ยงเบนมาตรฐาน σ_k ขององค์ประกอบที่ k^{th} จะสามารถตั้งให้เท่ากับค่า d โดยที่การคำนวณค่าของ a และ c จะใช้การคูณเวกเตอร์แบบ dot product ส่วนตัวอย่างพิกเซลที่ไม่ใช่พื้นหลังที่กำลังถูกพิจารณาจะถูกพิจารณาเป็นเงาที่กำลังเคลื่อนไหวถ้าค่าของ a อยู่ภายในค่าเบี่ยงเบนที่ 2.5 และ $\tau < c < 1$

2.3 การระบุตำแหน่งของวัตถุภายในภาพ (Object Location)

การระบุตำแหน่งของวัตถุในภาพที่เป็นตัวบุคคลในงานวิจัยนี้ จะตรวจจับตัวบุคคลจากภาพเคลื่อนไหวที่ได้จากการลบภาพออกจากแบบจำลองพื้นหลังที่ผสมผสานหลายเกาส์เซียน ซึ่งจะได้ภาพที่มีส่วนที่เคลื่อนไหว ซึ่งจะต้องนำมาระบุตำแหน่งของบุคคลในภาพ โดยมีขั้นตอนหลัก ๆ คือ การตรวจจับขอบของวัตถุ เพื่อจะใช้การตรวจจับเส้นขอบแสดงรูปร่างจาก Freeman Chain Code ซึ่งก็จะได้ตำแหน่งจุดที่ล้อมรอบวัตถุ จากนั้นจะใช้วิธีการประมาณค่าเฉพาะส่วนที่เป็นเหลี่ยมและมุม เพื่อให้เหลือจุดที่สำคัญของวัตถุ จากนั้นเราก็จะสามารถทราบคุณสมบัติอื่น ๆ ในเบื้องต้นของวัตถุได้ เช่น กรอบสี่เหลี่ยมของวัตถุ, จุดศูนย์กลางของวัตถุ, จุดศูนย์ถ่วงมวลของวัตถุ, ความกว้าง, ความสูง, พื้นที่ เป็นต้น ดังตัวอย่างในภาพประกอบที่ 2-8



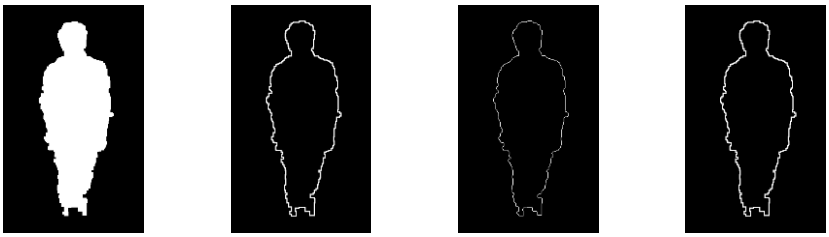
(ก) (ข) (ค) (ง) (จ)

ภาพประกอบที่ 2-8 ตัวอย่างขั้นตอนย่อยของการระบุตำแหน่งของวัตถุภายในภาพ (ก) ภาพต้นฉบับ (ข) ขอบของวัตถุ (ค) เส้นขอบแสดงรูปร่าง (Contour) (ง) ส่วนที่เป็นเหลี่ยมและมุม (จ) คุณสมบัติของวัตถุ เช่น กรอบสี่เหลี่ยม, จุดศูนย์กลาง, จุดศูนย์ถ่วงมวล, ความกว้าง, ความสูง, พื้นที่ เป็นต้น

2.3.1 การตรวจจับขอบของวัตถุ (Edge Detection)

ในงานวิจัยนี้จะใช้วิธีการตรวจจับขอบของวัตถุโดย Laplacian [87] โดยเป็นการหาอนุพันธ์อันดับสองเพื่อการตรวจจับขอบทั้งแกน X และ Y แล้วนำมารวมกันตามสมการที่ (2.7) ซึ่งขอบที่ได้จาก Laplacian จะค่อนข้างมีความหนาและชัดเจน เมื่อเทียบกับวิธีอื่น ๆ ตามภาพประกอบที่ 2-9

$$I_{Lp} = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \tag{2.7}$$

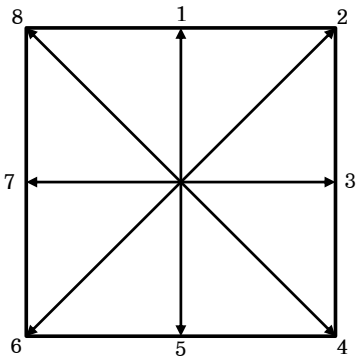


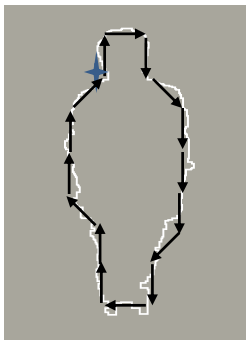
(ก) (ข) (ค) (ง)

ภาพประกอบที่ 2-9 ตัวอย่างการตรวจจับขอบของวัตถุ
(ก) ภาพต้นฉบับ (ข) ขอบของวัตถุที่ได้จากวิธีการ Sobel (ค) ขอบของวัตถุที่ได้จากวิธีการ Canny (ง) ขอบของวัตถุที่ได้จากวิธีการ Laplacian

2.3.2 การตรวจจับเส้นขอบแสดงรูปร่าง (Contour Approximation)

การตรวจจับเส้นขอบที่แสดงรูปร่างของวัตถุ จะเป็นการหาทิศทางของกรอบที่ล้อมรอบวัตถุแบบปิด โดยจะใช้การตรวจจับเส้นขอบแสดงรูปร่างจาก Freeman Chain Code [92] ซึ่งจะเป็นการหาทิศทางของขอบในวัตถุที่ปิด ซึ่งก็จะได้ตำแหน่งของเส้นขอบแสดงรูปร่างที่จะแสดงในรูปแบบของจุดที่เป็นคู่อันดับ x,y ซึ่งแสดงตัวอย่างการหาเส้นขอบแสดงรูปร่างจาก Freeman Chain Code ในภาพประกอบที่ (2-10)



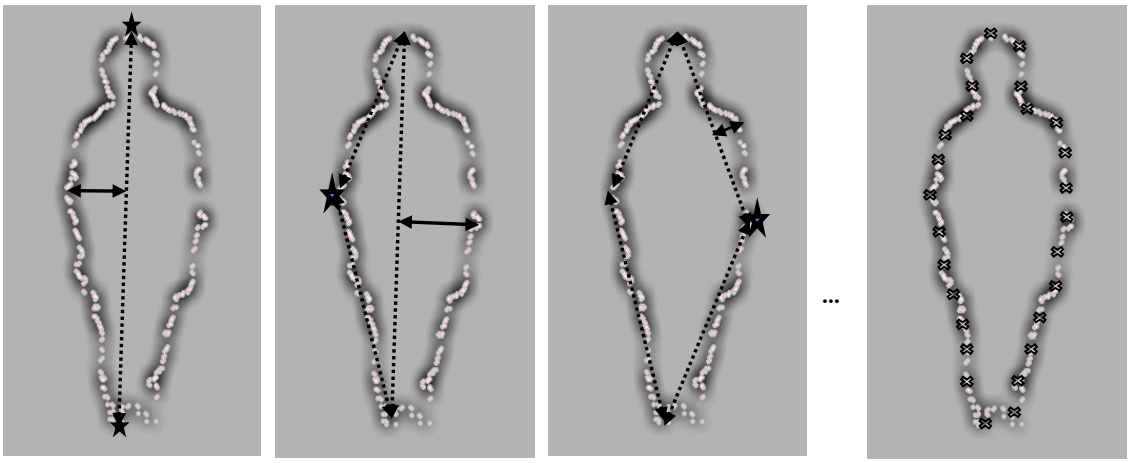


CC : 1-3-5-4-5-5-5-6-5-7-1-1-8-1-1-2

ภาพประกอบที่ 2-10 ตัวอย่างการตรวจจับเส้นขอบแสดงรูปร่างจาก Freeman Chain Code

2.3.3 การประมาณค่าเฉพาะส่วนที่เป็นเหลี่ยมและมุม (Polygon Approximation)

การประมาณค่าเฉพาะส่วนที่เป็นเหลี่ยมและมุมในงานวิจัยนี้ จะใช้ขั้นตอนวิธีของ Ramer-Douglas-Peucker [93] โดยขั้นตอนวิธีจะช่วยลดจำนวนของจุดในส่วนโค้งต่าง ๆที่ประกอบด้วยจุดจำนวนมาก ซึ่งในขั้นแรกขั้นตอนวิธีจะวัดระยะเพื่อหาจุดที่ไกลที่สุดหรือใช้วิธีการตั้งเป้าไว้ที่จุดแรกและจุดสุดท้าย ซึ่งจะได้ 2 Segment จากการลากเส้น จากนั้นให้หาระยะทางที่ตั้งฉากต่อจุดที่ไกลที่สุดจากเส้นที่ลากแบ่ง Segment (Perpendicular Distance) จากนั้นให้ลากเส้นแบ่ง Segment ที่จุดที่ไกลที่สุดนั้น แล้วทำซ้ำแบบเดิมไปเรื่อย ๆ เป็น Recursive ทุก ๆ Segment ที่แบ่ง โดยมีเงื่อนไขในการหยุดคือ หากค่าระยะทางที่ตั้งฉากต่อจุดที่ไกลที่สุดจากเส้นที่ลากแบ่ง Segment มีค่าน้อยกว่าค่าเอฟซิลอน (ε) เป็นเงื่อนไขในการหยุดการแบ่ง Segment ต่อไป ซึ่งได้แสดงตัวอย่างไว้ในภาพประกอบภาพประกอบที่ 2-11



ภาพประกอบที่ 2-11 ตัวอย่างการประมาณค่าเฉพาะส่วนที่เป็นเหลี่ยมและมุม

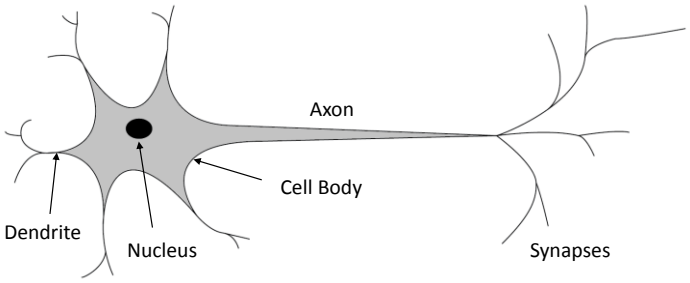
2.4 วิธีการจำแนกประเภทของข้อมูล (Classification Method)

2.4.1 โครงข่ายประสาทเทียม (Artificial Neural Network)

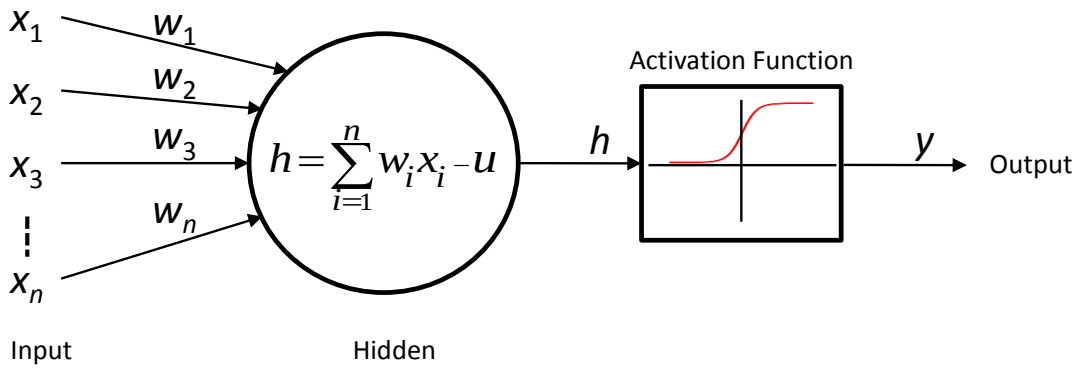
การประยุกต์ด้านระบบ อัจฉริยะที่เป็นปัญญาประดิษฐ์ หนึ่งในนั้นที่เป็นที่สนใจมากคือการประยุกต์โครงข่ายประสาททางชีวภาพ โดยนักวิจัยได้จำลองและออกแบบโครงข่ายประสาทเทียม (Artificial Neural Network) [94] เพื่อที่จะแก้ไขปัญหาและทำสิ่งต่าง ๆหลายอย่างแทนมนุษย์ เช่น การจดจำรูปแบบ (Pattern Recognition), การทำนาย (Prediction), การจำแนกประเภทของข้อมูล (Data Classification), การจัดกลุ่มข้อมูล (Data Clustering), การฟิตข้อมูล (Data Fitting) เป็นต้น ซึ่งในงานวิจัยนี้จะกล่าวถึงโครงข่ายประสาทเทียมแบบดั้งเดิม (Traditional Artificial Neural Network) ที่ประกอบด้วย Multi-layer Feed Forward และ Back-propagation Learning สำหรับ Multi-layer perceptron ซึ่งใช้ในการเรียนรู้เพื่อนำไปใช้ในการจดจำรูปแบบและการจำแนกประเภทของข้อมูลที่เป็นฟีเจอร์หนึ่งมิติ

โดยโครงข่ายประสาททางชีวภาพจะประกอบด้วย เซลล์ประสาท (Neuron) และเส้นประสานประสาท (Synapse) ซึ่งเกิดการเชื่อมโยงกันระหว่างเซลล์ประสาท ทำให้กลายเป็นระบบโครงข่ายที่ทำงานร่วมกัน ซึ่งโครงข่ายใช้กระแสประสาทที่เป็นไฟฟ้าเชื่อมโยงประสานกัน มีการคำนวณและส่งข้อมูลระหว่างกันสามารถเก็บความรู้ จดจำประสบการณ์ ในเซลล์ต่าง ๆ เพื่อนำความรู้ที่ได้การเรียนรู้มาไปใช้ในการวิเคราะห์ ตีความหรือคาดคะเนความหมายของข้อมูลที่ลักษณะใกล้เคียงกัน ซึ่งโครงข่ายงานประสาทเทียมก็เลียนแบบและจำลองวิธีการทำงานจากโครงข่ายประสาททางชีวภาพเช่นกัน

โดยโครงสร้างของโครงข่ายประสาทเทียม จะประกอบด้วย Input Layer ซึ่งได้รับจากการป้อนซึ่งอาจจะเป็น Output จากโหนดอื่นก็ได้, Hidden Layer ที่จะนำเอา Input มาคูณกับ Weight (W) ของแต่ละตัว แล้วนำไปรวมกันที่ Neuron ซึ่งจะมี Bias (u) เป็นตัวปรับค่า ซึ่งก็จะได้ผลลัพธ์จากโหนด (h) ตามสมการที่ (2.8) จากนั้นจะนำไปผ่าน Activation Function เพื่อบีบค่าให้อยู่ในช่วงที่ต้องการ จากนั้นก็จะส่งไปยังโหนดถัดไปเรื่อย ๆแล้วทำซ้ำแบบเดิม จนไปถึง Output Layer ซึ่งจะเป็นคำตอบที่ได้จาก Neuron Network (แสดงตัวอย่างตามภาพประกอบที่ 2-12) ซึ่งกระบวนการที่ได้กล่าวมาข้างต้นนั้น จะเรียกว่า Feed Forward ซึ่งจะเป็นการใช้งาน Neuron Network ที่ได้มีการเรียนรู้ค่ามาก่อนแล้ว ซึ่งสิ่งสำคัญคือต้องเรียนรู้มาก่อน เพื่อให้ได้ Output ที่มีความถูกต้องแม่นยำมากที่สุด



(ก) โครงข่ายประสาททางชีวภาพ



(ข) โครงข่ายประสาทเทียม

ภาพประกอบที่ 2-12 โครงข่ายประสาททางชีวภาพและโครงข่ายประสาทเทียม

$$h = \sum_{i=1}^{n} w_i x_i - u \qquad (2.8)$$

โดยกระบวนการเรียนรู้ของ Neuron Network จะใช้ Back-propagation Learning เพื่อเรียนรู้ค่า Weight (W) และค่า Bias (u) โดยมีการส่งตัวอย่างของข้อมูล ที่ประกอบไปด้วย Input และ Target (Ground tooth) ที่ได้จากการสังเกตหรือเก็บข้อมูลมา นำมาใส่ลงไปในโครงข่าย จากนั้นจะแพร่ค่าความผิดพลาดกลับมาเพื่อปรับปรุงค่าต่าง ๆในโครงข่าย ซึ่งสามารถอธิบายเป็นขั้นตอนวิธีตามภาพประกอบที่ 2-13

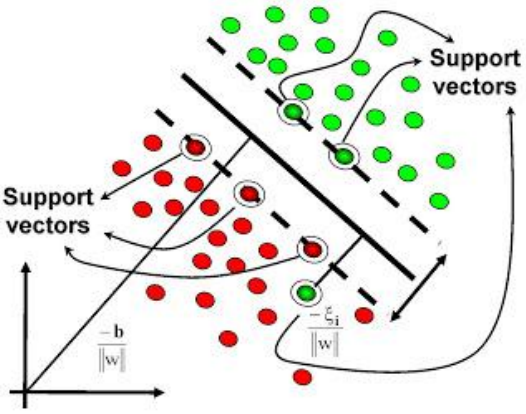
Back-propagation Algorithm

1. กำหนดค่าเริ่มต้นของ Weight โดยการสุ่มเป็นค่าน้อย ๆ
2. เลือกตัวอย่างข้อมูลแบบสุ่มมา 1 ค่า
3. แพร่ข้อมูลไปด้านหน้าแบบ Feed Forward เพื่อรอดู Output สุดท้าย ใน Output Layer
4. คำนวณค่า Error (δ_i^L) ใน Output โดยให้ $O_i = y_i^L$
   ตามสมการ $\delta_i^L = g'(h_i^L)\left[d_i^u - y_i^L \right]$
   โดยที่ h_i^L แทนค่ารวม net ที่ใส่ไปยัง หน่วยที่ i และ g' คือปริพันธ์จาก g ที่เป็น Activation Function
5. คำนวณค่า Delta สำหรับเลเยอร์ก่อนหน้านั้นโดยการแพร่กลับ Error ย้อนกลับไป
   ตามสมการ $\delta_i^l = g'(h_i^l)\sum_j W_{ij}^{l+1}\delta_j^{l+1}; \; for\, l = (L-1),...,1$
6. อัพเดทค่าน้ำหนักโดยใช้ $\Delta W_{ji}^l = \eta \delta_i^l y_j^{l-1}$ โดยที่ η คือค่าอัตราการเรียนรู้
7. ไปยังข้อที่ 2. แล้วทำซ้ำเรื่อย ๆ จนกว่าค่า Output ที่ออกมาจะได้ตามที่ตั้งค่าไว้

ภาพประกอบที่ 2-13 ขั้นตอนวิธีการแพร่ย้อนกลับสำหรับการเรียนรู้ของโครงข่ายประสาทเทียม

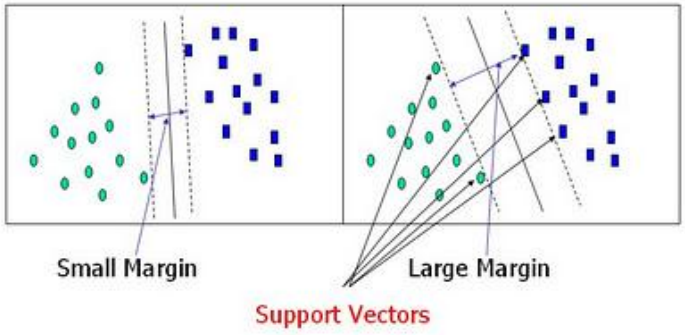
2.4.2 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

Support Vector Machine [95] เป็นขั้นตอนวิธีการเรียนรู้ของเครื่องแบบที่ต้องสอนข้อมูล (Supervised Machine Learning) ที่สามารถใช้ได้ทั้งการจำแนกประเภทของข้อมูลและการวิเคราะห์การถดถอย โดยที่มักจะถูกใช้ในการจำแนกประเภทของข้อมูลและรู้จำรูปแบบมากกว่า โดยวิธีการของ SVM จะ Plot ข้อมูลแต่ละตัวลงในระนาบ n มิติ โดยที่ n เป็นขนาดของฟีเจอร์ จากนั้นทำการหา Hyperplane ที่สามารถแยกกลุ่มของออกจากกันอย่างชัดเจน โดยใช้ฟังก์ชันหรือเรียกว่า Kernel ต่าง ๆที่เหมาะสมแต่ละกับข้อมูลตามความซับซ้อนที่ต่างกัน เช่น Linear, Sigmoidal, Polynomial, Radial เป็นต้น ที่จะสามารถแยกประเภทโดย Hyperplane ได้



ภาพประกอบที่ 2-14 ตัวอย่างการแยกข้อมูลโดย SVM ใน 2 มิติ

โดยการเลือกความเหมาะสมของตำแหน่งของการวาง Hyperplane นั้นจะต้องเลือกให้สามารถแบ่งข้อมูลได้อย่างชัดเจน และอยู่ห่างจากเวกเตอร์ของข้อมูลที่ถูก Plot ลงในระนาบให้มากที่สุด โดยใช้การหา Support Vector ที่ทำให้สามารถลากเส้น Hyperplane แล้วมี Margin ระหว่างข้อมูลมากที่สุดดังตัวอย่างในภาพประกอบที่ 2-15



ภาพประกอบที่ 2-15 ตัวอย่างการหาค่า Margin ของ Support Vector Machine [95]

2.5 สรุป

ในงานวิจัยนี้ภาพที่เข้ามาจะต้องผ่านกระบวนการประมวลก่อนในขั้นต้น เพื่อปรับปรุงภาพก่อน โดยในขั้นตอนแรกภาพความลึกที่ได้มาอาจจะมีหลุมที่ไม่มีค่าความลึก และมีค่าความแปรปรวนในบางบริเวณที่เป็นพื้นที่เล็ก ๆคล้ายกับสัญญาณรบกวนเกลือและพริกไทย จึงจำเป็นต้องใช้ตัวกรองสัญญาณแบบมัธยฐาน เพื่อปรับปรุงภาพในขั้นต้นก่อนกระบวนการตรวจจับบุคคลจากการจับภาพเคลื่อนไหว การลบภาพออกจากแบบจำลองพื้นหลังที่ผสมผสานหลายเกาส์เซียนและการประยุกต์การตรวจจับเงา เพื่อแก้ปัญหาต่าง ๆได้ เช่น ภาพพื้นหลังที่มีความซับซ้อน ภาพพื้นหลังที่มีการเปลี่ยนแปลงได้หลายรูปแบบ การเปลี่ยนแปลงของแสงเมื่อวัตถุเข้ามาในเฟรม เงาของวัตถุที่เคลื่อนไหวในภาพ ทำให้ค่าของจุดสีที่ในแบบจำลองภาพพื้นหลังเปลี่ยนแปลง

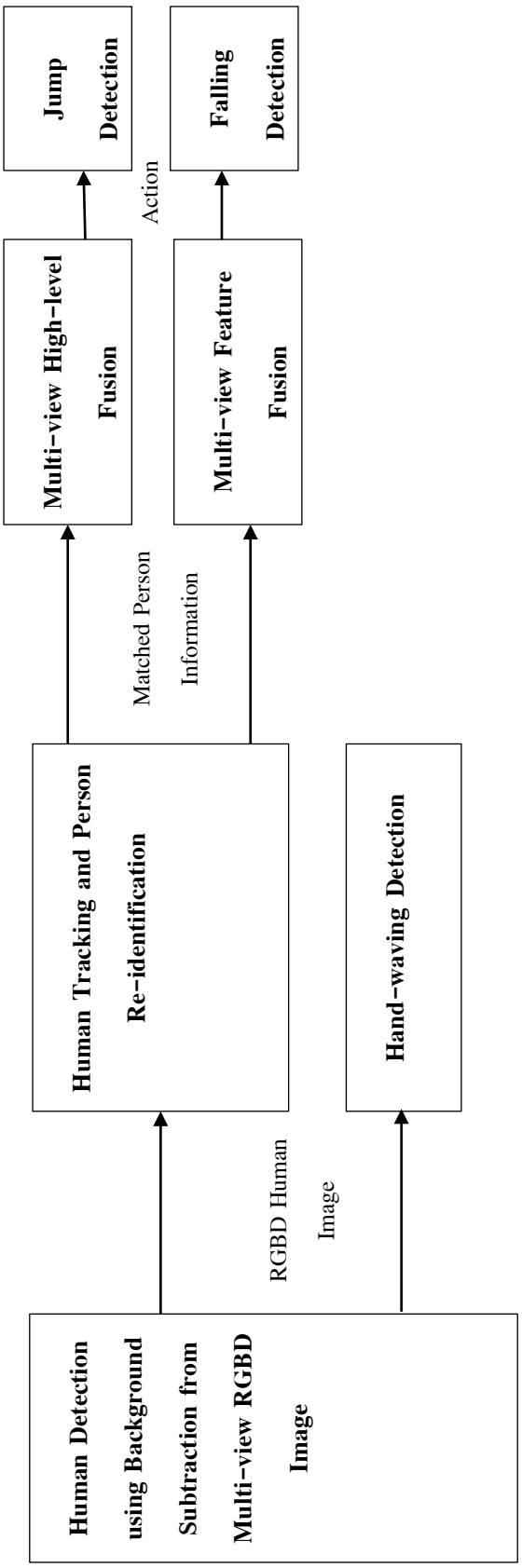
ไป เป็นต้น หลังจากนั้นจะใช้กระบวนการทางสัณฐานวิทยาของภาพ เพื่อลบสัญญาณรบกวน บางส่วนและเปลี่ยนแปลงรูปร่างหรือโครงสร้างของวัตถุในภาพ เมื่อภาพที่มีส่วนที่เคลื่อนไหวอยู่ในภาพซึ่งจะถูกสันนิษฐานว่าเป็นตัวบุคคล ก็จะใช้การระบุตำแหน่งของวัตถุภายในภาพ โดยมีขั้นตอนหลัก ๆ คือ การตรวจจับขอบของวัตถุ เพื่อจะใช้การตรวจจับเส้นขอบแสดงรูปร่างจาก Freeman Chain Code ซึ่งก็จะได้ตำแหน่งจุดที่ล้อมรอบวัตถุ จากนั้นจะใช้วิธีการประมาณค่าเฉพาะส่วนที่เป็นเหลี่ยมและมุม เพื่อให้เหลือจุดที่สำคัญของวัตถุ จากนั้นเราก็จะสามารถทราบคุณสมบัติอื่น ๆในเบื้องต้นของวัตถุได้ เช่น กรอบสี่เหลี่ยมของวัตถุ, จุดศูนย์กลางของวัตถุ, จุดศูนย์ถ่วงมวลของวัตถุ, ความกว้าง, ความสูง, พื้นที่ เป็นต้น โดยที่ตัวข้อมูลของบุคคลจะถูกนำไปแบบจำลองเพื่อหาฟีเจอร์ที่เป็นคุณลักษณะเด่นและจะใช้วิธีการจำแนกประเภทของข้อมูล ในการรู้จำรูปแบบของข้อมูล ซึ่งถูกสอนข้อมูลโดยการเรียนรู้จากกลุ่มตัวอย่างก่อนหน้าแล้ว

บทที่ 3
ระเบียบวิธีวิจัย

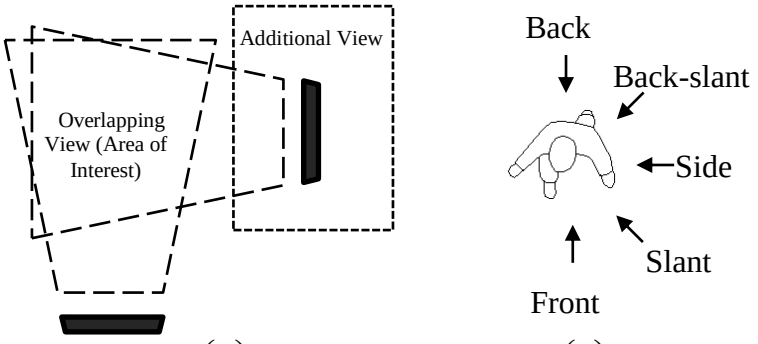
ในระเบียบวิธีวิจัยจะนำเสนอวิธีการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง (Multi-view Human Tracking and Person Re-identification) และในส่วนถัดมาจะเป็นส่วนของการรู้จำท่าทางจากหลายมุมมอง ที่ประกอบไปด้วยสองส่วนคือ การฟิวชันข้อมูลในระดับสูงจากหลายมุมมอง (Multi-view High-level Fusion for Action Recognition) และการฟิวชันฟีเจอร์ในระดับล่างจากหลายมุมมอง (Multi-view Feature Fusion for Action Recognition using Layer Fusion Model) ซึ่งจะใช้ข้อมูลจากการจับคู่บุคคลที่ตรงกันจากการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมองเพื่อนำฟีเจอร์ของบุคคลเดียวกันมาใช้ในการรู้จำท่าทางต่อไป หลังจากนั้นคำตอบที่เป็นท่าทางของบุคคลจะมีส่วนในการตัดสินใจของระบบตรวจจับการล้ม (Falling Detection) และระบบตรวจจับการกระโดด (Jump Detection) นอกจากนี้ยังมีระบบตรวจจับการโบกมือขอความช่วยเหลือ (Hand Waving Detection) ที่เป็นส่วนที่อิสระออกมาจากส่วนอื่น ซึ่งได้แสดงไว้ตามแผนผังของงานวิจัยโดยภาพรวมดังภาพประกอบที่ 3-1

3.1 การรู้จำท่าทางโดยการฟิวชันข้อมูลในระดับสูงจากหลายมุมมอง (Multi-view High-level Fusion for Action Recognition)

สำหรับในงานวิจัยนี้ได้ใช้การรู้จำท่าทางจากมุมมองเดี่ยว โดย P.Chawalitsittikul และคณะ [4] โดยข้อมูลภาพ RGB และความลึกจากหลายมุมมอง โดนมุมมองระหว่างกล้อง 2 กล้องตั้งฉากกัน โดยกล้องติดอยู่กับที่ไม่มีการเคลื่อนไหว ซึ่งทั้งสองต้องส่องไปยังบริเวณที่เดียวกัน ตามภาพประกอบที่ 3-2 (ก) โดยแนวความคิดของการฟิวชันข้อมูลในระดับคำตอบได้มาจากการที่สังเกตการณ์รู้จำท่าทางในหลากหลายมุมมองที่ส่องไปยังมนุษย์ จะพบว่าประสิทธิภาพในการรู้จำต่าง ๆขึ้นอยู่กับมุมมอง ตัวอย่างเช่น ท่านอนจะมีความถูกต้องสูงในมุมเฉียงด้านหน้า และด้านข้าง แต่จะมีความถูกต้องต่ำในด้านหน้าและด้านหลัง ซึ่งวิธีการของงานวิจัยนี้จะเพิ่มความถูกต้องของการรู้จำโดยสร้างฟังก์ชันวัดค่าความน่าเชื่อถือของคำตอบ (Weighting Function) โดยใช้มุมมองเป็นเกณฑ์ ซึ่งแบ่งเป็นมุม ด้านหน้า (Front), เฉียงด้านหน้า (Slant), ด้านข้าง (Side), เฉียงด้านหลัง (Back-slant), และด้านหลัง (Back) ตามภาพประกอบที่ 3-2 (ข)

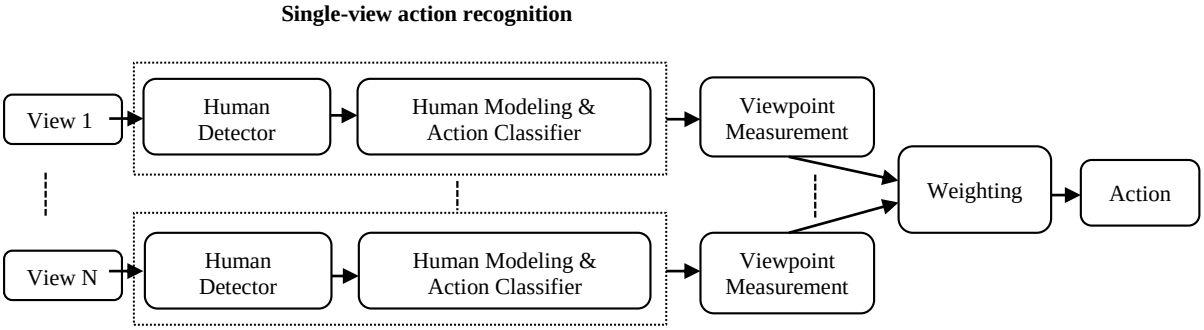


ภาพประกอบที่ 3-1 แผนผังของระบบต่าง ๆในงานวิจัยโดยภาพรวม



ภาพประกอบที่ 3-2 การตั้งมุมมองและการแบ่งมุมมองของระบบการฟิวชันข้อมูลในระดับคำตอบสำหรับการรู้จำท่าทางพื้นฐานของมนุษย์จากหลายมุมมอง

ซึ่งระบบจะแบ่งส่วนประกอบเป็น 3 ส่วน ส่วนแรกเป็นการรู้จำท่าทางในมุมมองเดี่ยวซึ่งประกอบด้วยการตรวจจับมนุษย์และการสร้างแบบจำลองมนุษย์ร่วมกันกับการรู้จำท่าทาง ส่วนที่สองเป็นส่วนของการตรวจวัดมุมมองเพื่อนำไปใช้ในฟังก์ชันวัดค่าความน่าเชื่อถือของคำตอบซึ่งเป็นส่วนที่ 3 ซึ่งจะได้คำตอบสุดท้ายโดยของการรู้จำท่าทางจากการฟิวชันข้อมูลในระดับคำตอบจากระบบตามภาพประกอบที่ 3-3



ภาพประกอบที่ 3-3 ภาพรวมของระบบการฟิวชันข้อมูลในระดับคำตอบสำหรับการรู้จำท่าทางพื้นฐานของมนุษย์จากหลายมุมมอง

3.1.1 การรู้จำท่าทางจากมุมมองเดี่ยว (Single-view Action Recognition)

ในส่วนนี้จะเป็นการอธิบายถึงการรู้จำท่าทางจากมุมมองเดี่ยว โดย P.Chawalitsittikul และคณะ [4] ซึ่งในขั้นตอนแรกจะต้องดึงมนุษย์ออกจากภาพความลึกโดยใช้วิธีการหักลบพื้นหลังแบบปรับเปลี่ยนค่าได้ตามสภาพ จากนั้นในขั้นตอนที่สองจะหาคุณสมบัติตำแหน่งของของศีรษะและขา ส่วนขั้นตอนสุดท้ายจะมีการนำโครงข่ายประสาทเทียม (Artificial Neural Network) มาใช้ในการแยกท่าทางทั้ง 5 ท่าทาง

(ก) การตรวจจับบุคคล (Human Detection)

ในการตรวจจับบุคคลจะใช้ตัวแยกพื้นหลังแบบปรับเปลี่ยนได้ โดยจะมีพื้นฐานของการจดจำพื้นหลังอยู่บนแบบจำลองเกาส์เซียนหลายแบบจำลองที่ผสมรวมกัน ซึ่งจะใช้ในการแยกระหว่างฉากหน้าที่เคลื่อนไหวและพื้นหลัง โดยจะถูกนำมาใช้งานเพื่อตรวจจับบริเวณที่มีการเคลื่อนไหวซึ่งจะกำหนดว่าบริเวณนั้นคือบุคคล โดยวิธีการนี้จะทนต่อการเปลี่ยนแปลงที่ไม่คงที่ของแบบจำลองพื้นหลังซึ่งถูกกำหนดโดยแบบจำลองเกาส์เซียน 3-5 รูปแบบ หลังจากนั้นบริเวณที่เป็นตัวบุคคลจะถูกเติมด้วยข้อมูลความลึกและสีในเส้นขอบของวัตถุ โดยเส้นขอบที่เป็นสีคือบริเวณที่ซ้อนทับกันของภาพสี ($\boldsymbol{I_C}$) และขอบของวัตถุในภาพความลึก ($\boldsymbol{O_D} - E(\boldsymbol{O_D})$ โดยเส้นขอบจะถูกกำหนดโดยการลบกันของก้อนวัตถุความลึกที่เป็นบุคคล และ erosion (E) of $\boldsymbol{O_D}$, ดังที่ได้แสดงไว้ในสมการที่ (3.1)

$$\mathbf{H}_{\mathrm{DC}} = \mathrm{E}(\mathbf{I}_{\mathrm{D}}) \cup (\mathbf{I}_{\mathrm{C}} \cap (\mathbf{O}_{\mathrm{D}} - \mathrm{E}(\mathbf{O}_{\mathrm{D}}))) \tag{3.1}$$

(ข) แบบจำลองโครงสร้างของตัวบุคคลและวิธีการรู้จำ (Human Modeling and Classification Method)

ในการสร้างแบบจำลองของตัวบุคคล กระบวนการจะเริ่มต้นที่การหา Center of Mass $(\bar{\mathrm{x}}, \bar{\mathrm{y}})$ ใน Human Object โดยอ้างอิงตำแหน่งของหัวและขา ต่อจากนั้น จะค้นหา Vectors ที่มีขนาดมากที่สุด จากศูนย์กลางไปยังขอบของวัตถุในแต่ละ Quadrant โดย Vector (Δ_{i}) ทั้งสี่จะถูกรวมกันเป็นสอง Vector ซึ่งเป็น Vector ที่ชี้ไปยังหัวและขาตามลำดับ โดยที่พิจารณามุมที่กระทำกันน้อยที่สุดระหว่างสอง Vector ใด ๆ การรวมกันของ Vector จะถูกถ่วงน้ำหนักโดยค่า Magnitude ($\mathrm{D_v}$) และ Color Distance ($\mathrm{D_c}$) ของ Vector ทั้งสอง ตามสมการที่ (3.2) และ (3.3) โดย Color Distance ($\mathrm{D_c}$) สำหรับเฟรมถัดไป ($\mathrm{Cr_{t+1}}$) จะถูกอัพเดทค่าจากเฟรมปัจจุบันโดยสมการที่ (3.4) และค่าถ่วงค่าน้ำหนักของ First Vector (ω_1) สำหรับ Combining จะถูกแสดงไว้ในสมการที่ (3.5) และ (3.6)

$$\mathrm{D_{v_i}} = \sqrt{(\Delta_{\mathrm{x_i}} - \bar{\mathrm{x}})^2 + (\Delta_{\mathrm{y_i}} - \bar{\mathrm{y}})^2} \mid \mathrm{i} = 1 \ldots 4 \tag{3.2}$$

$$\mathrm{D_{c_i}} = \sqrt{(\Delta_{\mathrm{c_i}} - \mathrm{Cr_t})^2} \mid \mathrm{i} = 1 \ldots 4 \tag{3.3}$$

$$\mathrm{Cr_{t+1}} = \alpha_{\mathrm{c}}\mathrm{Cr_t} - (1 - \alpha_{\mathrm{c}})(\omega_1 \mathrm{C}_1 + \omega_2 \mathrm{C}_2) \tag{3.4}$$

$$\omega_{v_1} = \left[\sum_{j=1}^{2}\left(\frac{D_{v_2}}{D_{v_j}}\right)^{\frac{2}{m-1}}\right]^{-1} \quad (3.5)$$

$$\omega_{c_1} = \left[\sum_{j=1}^{2}\left(\frac{D_{c_2}}{D_{c_j}}\right)^{\frac{2}{m-1}}\right]^{-1} \quad (3.6)$$

การถ่วงค่าน้ำหนักจะเป็นการปรับค่าโดย Magnitude (D_v) และ Color Distance (D_c) โดยใช้ค่าอัลฟ่า (α_v) ดังที่ได้แสดงไว้ในสมการที่ (3.7) จากนั้น Second Vector (ω_2) จะถูกกำหนด โดย Complement ของ ω_1 ดังสมการที่ (3.8)

$$\omega_1 = \alpha_v\omega_v + (1-\alpha_v)\omega_c \quad (3.7)$$

$$\omega_2 = 1 - \omega_1 \quad (3.8)$$

สุดท้าย ตำแหน่งของ Vector ทั้งสองจะถูกรวมกันเป็น Vector เดียวกัน ก็จะได้ Vector ที่ชี้ไปยังขา $\vec{v}_l$ และ Vector ที่ชี้ไปยังหัว $\vec{v}_h$ โดยใช้การถ่วงน้ำหนักของแต่ละ Vector ตามสมการที่ (3.9)

$$P(x,y) = \omega_1 P_1(\Delta_{x_1}, \Delta_{y_1}) + \omega_2 P_2(\Delta_{x_2}, \Delta_{y_2}) \quad (3.9)$$

ฟีเจอร์ในแบบจำลองมนุษย์จะถูกกำหนดโดยใช้ Upper Vector (Center to Head $\vec{v}_h$) และ lower vector (Center to Legs $\vec{v}_l$), ซึ่งจะประกอบด้วย มุมระหว่างตัวและหัว (θ_h) ของ $\vec{v}_h$, มุมระหว่างตัวและขา (θ_l) ของ $\vec{v}_l$, ค่าความต่างกันของความลึก ระหว่างหัวและส่วนกลางที่ Center of Mass (D_h) และค่าความต่างกันของความลึก ระหว่างขาและส่วนกลางที่ Center of Mass (D_l) ซึ่งฟีเจอร์เหล่านี้ จะถูกนำไปใช้ในการรู้จำท่าทางโดยโครงข่ายประสาทเทียม (Artificial Neural Network) ที่ใช้ 1 เลเยอร์ซ่อน - 30 โหนด ด้วย Sigmoid Function ซึ่งจะถูกเรียนรู้จากกระบวนการของกระบวนการแพร่ย้อนกลับ (Back-propagation)

3.1.2 การฟิวชันท่าทางของมนุษย์ในระดับสูงจากหลายมุมมอง (High-level Multi-view Action Fusion)

แนวคิดการฟิวชันในระดับสูงของงานวิจัยนี้มาจากพื้นฐานเทคนิคการถ่วงน้ำหนักเพื่อจะหาว่าคำตอบจากมุมมองใดที่น่าเชื่อถือที่สุด โดยคำตอบของท่าทางที่มีค่าถ่วงน้ำหนักที่สุดจะถูกนิยามว่ามีความถูกต้องแม่นยำมากที่สุด การได้มาซึ่งวิธีการฟิวชันนี้ได้มาจาก Empirical Observations

ของการรู้จำท่าทาง โดยใช้มุมมองเดียวจากหลาย ๆ มุมที่ต่างกัน ซึ่งได้แสดงไว้ในภาพประกอบที่ 3-2

(ก) การทดลองเชิงประจักษ์ในการรู้จำท่าทางในมุมมองต่าง ๆ (Empirical Observation for Viewpoint Action Recognition)

ระหว่างที่ทำการทดลองมีข้อสังเกตว่าความถูกต้องของการรู้จำท่าทางในมุมมองเดี่ยวจะขึ้นอยู่กับมุมมองโดยที่ฟีเจอร์ของมนุษย์ที่ได้จากแต่ละมุมมอง จะเปลี่ยนแปลงไปตามสิ่งรบกวน (Noise), การบิดงอของมุมมอง (Perspective Distortion), หรือการขาดข้อมูลที่สมบูรณ์ (Lack of Information) อย่างเช่น ในท่านั่งจากมุมมองด้านหน้าจะมีความถูกต้องน้อยกว่ามุมมองด้านข้าง, ท่าก้มจากด้านหน้า เฉียงด้านหลังและด้านหลัง จะมีความถูกต้องน้อยกว่ามุมมองด้านข้าง ดังนั้น การทดลองรู้จำท่าทางของผู้วิจัยที่ได้กล่าวถึงไว้ในหัวข้อที่ 3.1.1 จะถูกแบ่งได้เป็น 5 มุมมอง เพื่อที่จะหาความถูกต้องในแต่ละมุมมองของทุกท่าทาง ซึ่งผลลัพธ์ได้แสดงไว้ในตารางที่ 3-1 ซึ่งแสดงความถูกต้องในท่าทางและมุมมองต่าง ๆ และตารางที่ 3-2 ได้แสดง Confusion Matrix ของท่าทางต่าง ๆจากทุกมุมมองโดยที่จะนำไปใช้ในกระบวนการฟิวชันข้อมูลในระดับคำตอบต่อไป

ตารางที่ 3-1 ความถูกต้องของการรู้จำในท่าทางและมุมมองต่าง ๆในมุมมองเดี่ยว

ท่าทาง	ความถูกต้องตามมุมมอง (%)					
	หน้า	เฉียงทางหน้า	ข้าง	เฉียงทางหลัง	หลัง	ค่าเฉลี่ย
ยืน / เดิน	81.21	76.59	75.93	72.86	61.83	73.68
นั่ง	85.61	44.75	53.72	29.97	93.00	61.41
ก้ม	62.57	73.61	85.80	3.95	6.32	46.45
นอน	0	78.67	98.06	89.56	0	53.25

ตารางที่ 3-2 Confusion Matrix ของท่าทางต่าง ๆจากทุกมุมมอง

		Target Actions (%)			
		ยืน / เดิน	นั่ง	ก้ม	นอน
Desired Actions (%)	ยืน / เดิน	74.14	23.75	32.27	27.06
	นั่ง	23.22	61.90	18.69	12.54
	ก้ม	2.46	14.03	48.52	7.75
	นอน	0.18	0.32	0.52	52.64

(ข) การวัดหาค่ามุมมอง (Viewpoint Measurement)

การวัดค่าหามุมมองเป็นการที่จะหาว่ามุมมองที่เป็นทิศทางที่กล้องชี้ไปหามนุษย์ที่กำลังติดตามท่าทางอยู่ ซึ่งการพิจารณาเพื่อจะหามุมมองจะถูกแบ่งเป็น 5 มุมมอง ประกอบไปด้วย ด้านหน้า เฉียงด้านหน้า ด้านข้าง เฉียงด้านหลัง ด้านหลัง ซึ่งการวัดหาค่ามุมมองนี้มีความสำคัญมากต่อการหาความถูกต้องของท่าทางซึ่งขึ้นอยู่กับมุมมอง โดยที่กระบวนการแรกของการหาฟีเจอร์ คือการตัดแบ่ง ซึ่งฟีเจอร์ที่จะใช้ในการรู้จำมุมมองซึ่งชี้ไปทางมนุษย์ได้แก่ความกว้าง แกน และความสูงโดยจะแบ่งตำแหน่งของมนุษย์เพื่อที่จะคำนวณฟีเจอร์ตาม แนวราบและแนวดิ่งดังแสดงไว้ในภาพประกอบที่ 3-4 (ก) และ 3-4 (ข) สำหรับฟีเจอร์ความกว้าง (Width) และแกนกลางของตัว (Axis) ดังแสดงไว้ในภาพประกอบที่ 3-4 (ค) โดยจะแบ่งวัตถุที่เป็นตัวมนุษย์เป็นส่วนที่เท่า ๆกันทั้งในแนวตั้งตามร่างกายจากด้านบนไปยังด้านล่างตามสมการที่ (3.10) สำหรับในแต่ละส่วน (Segment) ของฟีเจอร์ความกว้าง ($WF(i)$) นั้น สามารถหาค่าได้จากการโปรเจคชั่นค่าพิกเซลที่มีความสว่าง (I=1) จากภาพไบนารี่ของภาพความลึกในส่วนของวัตถุที่เป็นตัวมนุษย์ตามสมการที่ (3.11) และในส่วนของค่าแกนกลางของตัว ($AF(i)$) สามารถคำนวณโดยการหาตำแหน่งพิกเซลแรกที่มีความสว่าง (LP_x) และตำแหน่งสุดท้ายที่มีความสว่าง (RP_x) ในแต่ละส่วนเพื่อที่จะหาค่าของแกนกลางของตัวตามสมการที่ (3.12) และในส่วนของฟีเจอร์ความสูง ($HF(i)$) จะใช้การแบ่งส่วนตามแนวดิ่งตามร่างกายจากซ้ายไปขวาที่มีขนาดเท่ากันตามสมการที่ (3.13) จากนั้นค่าความสูงจะถูกคำนวณจากการโปรเจคชั่นค่าพิกเซลที่มีความสว่าง (I=1) เช่นเดียวกับค่าความกว้าง ตามสมการที่ (3.14) โดยที่ $rows$ คือความสูงของทั้งตัววัตถุ และ $cols$ คือความกว้างของทั้งตัววัตถุ ส่วน $PyWF(i), PyAF(i)$ คือ ตำแหน่งของ Segment ที่ i ของฟีเจอร์ความกว้างและแกน ในส่วนของ $PxHF(i)$ คือ ตำแหน่งของ Segment ที่ i ของฟีเจอร์ความสูง โดน i คือดัชนีลำดับของ Segment ที่ 1 ถึง l_w และ l_h ซึ่งเป็นขนาดของ Segment

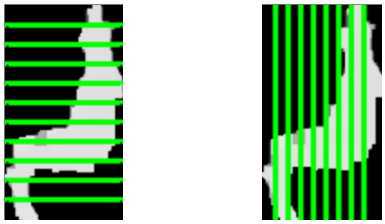
$$PyWF(i), PyAF(i) = (\frac{rows}{l_w} * i) \mid i = 1 \dots l_w \tag{3.10}$$

$$WF(i) = \sum_{x=1}^{cols} f\big(\mathrm{x}, PyWF(\mathrm{i})\big) \mid if(I == 1) \tag{3.11}$$

$$AF(i) = LP_{\mathrm{x}}\big(PyAF(i)\big) + \left(\frac{RP_{\mathrm{x}}\big(PyAF(i)\big) - LP_{\mathrm{x}}\big(PyAF(i)\big)}{2}\right) \tag{3.12}$$

$$PxHF(i) = (\frac{cols}{l_h} * i) \mid i = 1 \dots l_h \tag{3.13}$$

$$HF(i) = \sum_{y=1}^{rows} \mathrm{f}(PxHF_{\mathrm{i}}, \mathrm{y}) \mid i = 1 \dots l_h \mid if(I == 1) \tag{3.14}$$



(ก) การแบ่งส่วนในแนวราบ (ข) การแบ่งส่วนในแนวดิ่ง

(ค) ฟีเจอร์ความกว้าง (Width), แกนกลางของตัว (Axis) และ
ความสูง (Height)

ภาพประกอบที่ 3-4 ตัวอย่างการแบ่งส่วนและฟีเจอร์ในการวัดหาค่ามุมมอง

ในขั้นตอนสุดท้าย ฟีเจอร์ทั้งสามตัวจะถูกเรียงต่อกันเป็นเวกเตอร์ฟีเจอร์ เพื่อที่จะนำไปเรียนรู้โดยใช้วิธีการแยกประเภท เพื่อนำไปใช้ทดสอบการรู้จำมุมมอง โดยผู้วิจัยได้เลือกโครงข่ายประสาทเทียมที่ใช้ 1 เลเยอร์ซ่อน - 30 โหนด ด้วย Sigmoid Function เป็นวิธีในการแยกหมวดหมู่ของการวัดค่าหามุมมอง โดยลักษณะของการเรียงต่อกันเป็นเวกเตอร์ฟีเจอร์จะมีดังนี้

$$\begin{matrix}\{WF(1), WF(2), WF(3), \dots, WF(l_w),\\ AF(1), AF(2), AF(3), \dots, AF(l_a),\\ HF(1), HF(2), HF(3), \dots, HF(l_h)\}\end{matrix} \quad (3.15)$$

(ค) การฟิวชันข้อมูลระดับสูง (High-level Fusion Model)

สำหรับในขั้นตอนนี้จะอธิบายแบบจำลองการฟิวชันข้อมูลระดับสูง ซึ่งจะนำคำตอบของท่าทางในแต่ละมุมมองเดี่ยว โดยแบบจำลองจะสร้างฟังก์ชันในการถ่วงน้ำหนักที่ถูกกำหนดโดยอัตราความถูกต้องของการรู้จำท่าทางในแต่ละมุมมองตามที่ได้กล่าวในหัวข้อการทดลองเชิงประจักษ์ในการรู้จำท่าทางในมุมมองต่าง ๆ ซึ่งได้แสดงไว้ดังในตารางที่ 3-1 และ 3-2

โดยค่าความถูกต้องของแต่ละท่าทาง (A_{a_i,v_k}) ที่ได้จากการทดสอบในแต่ละมุมมองต่าง ๆ ดังที่แสดงไว้ตามตารางที่ 3-1 จะถูกนำกำหนดตามสมการที่ (3.16) ซึ่งสามารถหาค่าต่าง ๆโดยการเข้าถึงค่าที่อยู่ในตาราง ตามแถวซึ่งเป็นท่าทางและคอลัมน์ซึ่งเป็นมุมมอง

$$A_{a_i,v_k} = ta[v_k][a_i] \quad (3.16)$$

โดยที่ ta คือ ตารางที่ 3-1 ซึ่งความถูกต้องของการรู้จำในท่าทางที่ a_i และมุมมองต่าง ๆที่ v_k

สำหรับในแบบจำลองฟิวชันแบบเบื้องต้น ผู้วิจัยจะใช้ค่าความถูกต้องของแต่ละท่าทาง (A_{a_i,v_k}) ในการโหวตเพื่อเลือกคำตอบของท่าทางที่ดีที่สุด อย่างเช่น คำตอบท่าทางที่ได้จากมุมมองเดี่ยวสองมุมเป็นท่านั่ง และท่ายืน จากด้านหน้าและด้านข้าง ซึ่งค่าของ A_{a_i,v_k} จะเท่ากับ 85.61 และ 75.93 ตามลำดับ ดังนั้นท่านั่งซึ่งมีค่าความถูกต้องมากกว่าจะถูกเลือกเป็นคำตอบสุดท้ายของแบบจำลองฟิวชันแบบเบื้องต้น

อย่างไรก็ตามผู้วิจัยพบว่าในท่านั่งและท่านอนในบางมุมมองยังมีค่าความแม่นยำจากการฟิวชันที่ค่อนข้างน้อยอยู่ ผู้วิจัยจึงได้เสนอการใช้ Confusion Matrix ของท่าทางต่าง ๆจากทุกมุมมอง

สำหรับแบบจำลองฟิวชันแบบซับซ้อนจะเพิ่มพารามิเตอร์สองตัวจาก Confusion Matrix ได้แก่ ค่าความถูกต้องของท่าทางที่เป็นคำตอบที่กำลังพิจารณาค่าความน่าเชื่อถืออยู่ซึ่งเป็นค่า True Positive จากทุก ๆมุมมอง ($C_{a_i^t,a_i^d}$) และค่าความผิดพลาดของท่าทางที่กำลังพิจารณาค่าความน่าเชื่อถืออยู่ที่ตอบไปเป็นท่าทางที่เป็นคู่แข่งที่ได้จากมุมมองอื่น ($C_{a_i^t,a_j^d}$) ซึ่งค่าสองค่านี้จะได้จากการเข้าถึงค่าที่อยู่ในตาราง ตามแถวซึ่งเป็นความถูกต้องของท่าทางที่ได้จากการรู้จำและคอลัมน์เป้าหมายที่เป็นคำตอบจริง ๆ ตามสมการที่ (3.17)

$$C_{a_i^t,a_i^d} = tc[a_i^t][a_i^d] \tag{3.17}$$

โดยที่ tc คือ ตารางที่ 3-2 ที่เป็น Confusion Matrix ของท่าทางต่าง ๆจากทุกมุมมอง และ a_i^t และa_i^d เป็นความถูกต้องของท่าทางที่ได้จากการรู้จำและท่าทางที่เป้าหมายที่เป็นคำตอบจริง ๆ

ค่าความน่าเชื่อถือแบบจำลองฟิวชันซับซ้อนจะถูกกำหนดโดย $F_{a_i a_j,v_k}$ ณ ท่าทาง a_i กำลังพิจารณาอยู่ในมุมมองที่ v_k ด้วยท่าทางของอีกมุมมอง (Penalty Action) ที่ a_j ดังนี้:

$$F_{a_i a_j,v_k} = C_{a_i^t,a_i^d} + \alpha A_{a_i,v_k} - C_{a_i^t,a_j^d} \tag{3.18}$$

โดยที่ α คือ ค่าที่สามารถปรับค่าได้ตามความต้องการในแบบจำลองเพื่อให้ ความถูกต้องของแต่ละท่าทางของแต่ละมุมมอง (A_{a_i,v_k}) มีบทบาทน้อยหรือมาก

อย่างเช่น มีท่านั่งและยืน จากด้านหน้าและด้านข้าง ก็จะมี ท่าทางของอีกมุมมอง (Penalty Action) ได้แก่ ยืนและนั่ง ตามลำดับ ซึ่งจะต้องคำนวณค่าความน่าเชื่อถือแต่ละมุมมองแยกกัน

โดยมุมมองแรกที่เป็นท่านั่งจากด้านหน้า ค่าความน่าเชื่อถือ $F_{a_i a_j,v_k}$ จะคำนวณได้ตามตัวอย่างดังนี้

$$F_{a_{sitting} a_{stand},v_{front}} = 61.90 + \alpha 85.61 - 23.75$$

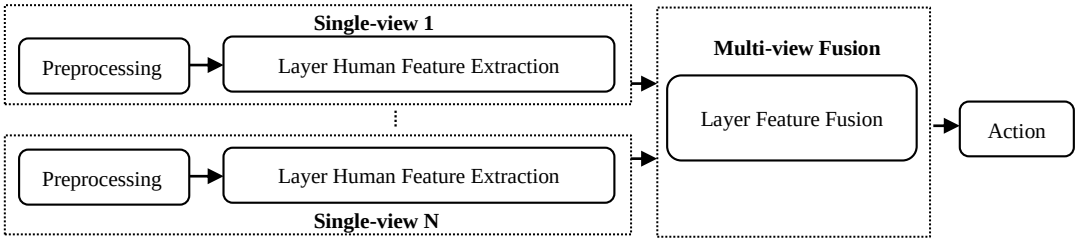
โดยมุมมองสองที่เป็นท่ายืนจากด้านข้าง ค่าความน่าเชื่อถือ $F_{a_i a_j,v_k}$ จะคำนวณได้ตามตัวอย่างดังนี้

$$F_{a_{stand}a_{sitting},v_{side}} = 74.14 + \alpha 75.93 - 23.22$$

และเมื่อคำนวณเสร็จทั้งสองก็จะต้องเลือกท่าทางจากมุมมองที่คำนวณแล้วค่าความน่าเชื่อถือ($F_{a_i a_j,v_k}$) มากกว่าจะถูกเลือกเป็นคำตอบสุดท้ายของแบบจำลองฟิวชันแบบซับซ้อน

3.2 การรู้จำท่าทางโดยการฟิวชันฟีเจอร์ในระดับล่างจากหลายมุมมอง (Multi-view Feature Fusion for Action Recognition using Layer Fusion Model)

การรู้จำท่าทางโดยการฟิวชันฟีเจอร์ในระดับล่างจากหลายมุมมอง จะใช้แบบจำลองแบบเลเยอร์แบบแนวตั้ง ซึ่งแบ่งพื้นที่เท่า ๆกันเพื่อถอดข้อมูลคุณลักษณะต่าง ๆในแต่ละเลเยอร์ เพื่อใช้ในการฟิวชันจากหลายมุมมอง โดยวิธีการนี้จะจัดอยู่ในฟีเจอร์ประเภทกริด ซึ่งมีความซับซ้อนในการสร้างแบบจำลองน้อย แต่ต้องแลกมากับการได้ข้อมูลที่ซ้ำกัน หรือข้อมูลที่ไม่มีนัยสำคัญในบางตำแหน่ง ซึ่งโมดูลหลักจะประกอบไปด้วย 1. กระบวนการเตรียม (Preprocessing) สำหรับปรับปรุงคุณภาพของภาพ 2. การทำแบบจำลองมนุษย์และการสกัดฟีเจอร์โดยใช้โมดูลสกัดฟีเจอร์แบบเลเยอร์จากมุมมองเดี่ยว (Layer Human Feature Extraction) 3. การฟิวชันกันของฟีเจอร์แบบเลเยอร์ (Layer Feature Fusion) ของมุมมองเดี่ยวจากทุกมุมมองลงในแบบจำลองเดียวกัน และจะถูกนำไปแยกแยะ (Action Classification) เป็นท่าทาง ซึ่งได้แสดงภาพรวมของระบบไว้ในภาพประกอบที่ 3-5



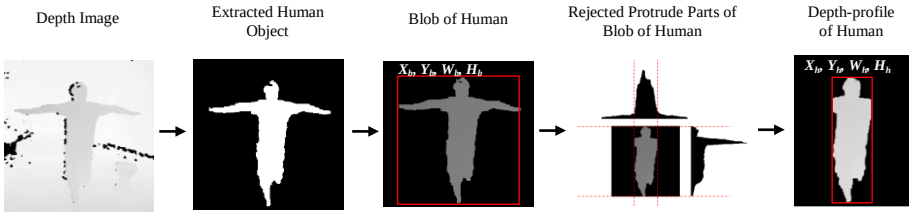
ภาพประกอบที่ 3-5 ภาพรวมของระบบรู้จำท่าทางของมนุษย์ที่ใช้แบบจำลองเลเยอร์เพื่อฟิวชันข้อมูลภาพสีและความลึกจากหลายมุมมอง

3.2.1 กระบวนการเตรียม (Preprocessing)

วัตถุประสงค์ของกระบวนการเตรียมคือ การปรับปรุงภาพความลึกและการแยกโครงสร้างของมนุษย์ที่เป็นตัวบุคคลเดี่ยว ๆ ออกจากพื้นหลังและตัดส่วนที่ยื่นออกมาที่ไม่จำเป็นสำหรับการรู้จำท่าทางพื้นฐานออกก่อนจะเข้าสู่กระบวนการสกัดฟีเจอร์ซึ่งได้อธิบายขั้นตอนตามภาพประกอบที่ 3-6 ในการทดลองของผู้วิจัย ซึ่งการปรับปรุงภาพความลึกจะทำโดยใช้ตัวกรองแบบมัธยฐานและดึงโครงสร้างของมนุษย์ที่อยู่ในฉากหน้าเป็นส่วนที่จะถูกตัดออกมาจากพื้นหลังโดยใช้การตรวจจับการเคลื่อนไหวที่มีพื้นฐานมาจากการลบภาพออกจากแบบจำลองที่ผสมผสานเกาส์เซียน โดยภาพเคลื่อนไหวที่ถูกสกัดออกมา (I_m) แสดงตามภาพประกอบที่ 3-6 (ข) จากภาพความลึก (I_d) ในภาพประกอบที่ 3-6 (ก) โดยตั้งสมมุติฐานว่าสิ่งที่เคลื่อนไหวในภาพเป็นตัวบุคคลที่กำลัง

ตรวจจับอยู่ แต่ภาพเคลื่อนไหวที่ถูกสกัดออกมายังคงมีสิ่งรบกวนจากการตรวจจับการเคลื่อนไหว ซึ่งต้องมีกระบวนการที่จะมาลดลงโดยกระบวนการทางสัณฐานวิทยาโดยการเปิดและตามด้วยการปิด หลังจากนั้นภาพเคลื่อนไหวที่ถูกตัดออกมานั้น จะถูกเติมค่าความลึกของตัวมันเองภายในวัตถุในส่วนที่เป็นมนุษย์ จะได้เป็นภาพของบุคคลที่ถูกปรับปรุงแล้ว ซึ่งจะแทนด้วยคำว่าวัตถุ ($\boldsymbol{I_{mo}}$) ซึ่งจะได้มาจากการผสานส่วนที่เหมือนกัน (Intersecting) ระหว่าง $\boldsymbol{I_d}$ และ $\boldsymbol{I_m}$ โดยใช้ แอนโอเปอเรเตอร์ (AND Operation) : $\boldsymbol{I_{mo} = I_m \& I_d}$.

หลังจากนั้นกลุ่มก้อนของวัตถุจะถูกดึงออกมาอยู่ในลักษณะพิกัดของกรอบสี่เหลี่ยม (Bounding Rectangle) ที่ประกอบไปด้วยตำแหน่งบนซ้าย X_h, Y_h, ความกว้าง W_h, และความสูง H_h โดยใช้การระบุตำแหน่งจากเทคนิคการประมาณการขอบภาพ ดังที่แสดงไว้ในภาพที่ 3-6 (ค) อย่างไรก็ตามเทคนิคของผู้วิจัยมุ่งเน้นเฉพาะท่าทางที่เป็นพื้นฐานของมนุษย์ซึ่งไม่พิจารณาท่าทางที่เกิดจากมือและแขน จึงจำเป็นต้องตัดออกโดยใช้เทคนิคการวิเคราะห์ภาพฉายที่เป็นไบนารี่ (Image Binary Projection) ซึ่งแสดงตัวอย่างไว้ในภาพประกอบที่ 3-6 (ง) ซึ่งจะได้โครงสร้างของมนุษย์ในรูปแบบของความลึก ($\boldsymbol{I_{mo}}$) ซึ่งแสดงไว้ในภาพประกอบที่ 3-6 (จ) ซึ่งจะถูกใช้ในกระบวนการของการสกัดฟีเจอร์ต่อไป



(ก) (ข) (ค) (ง) (จ)

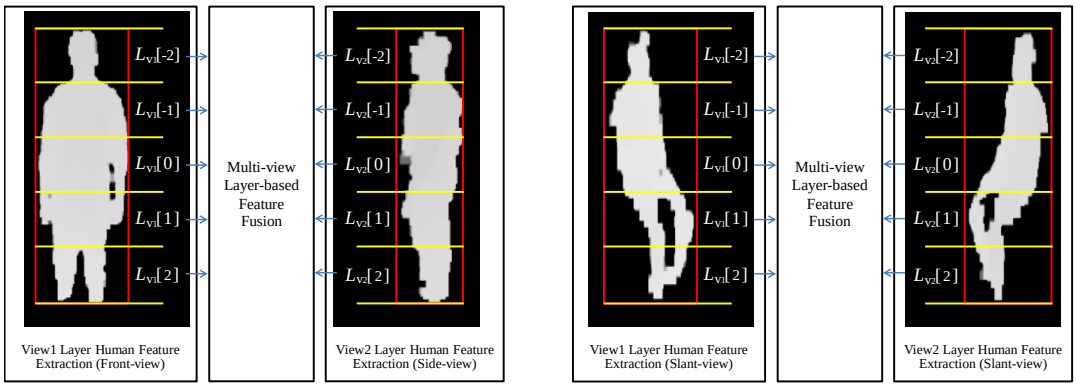
ภาพประกอบที่ 3-6 กระบวนการเตรียม (Preprocessing) สำหรับปรับปรุงคุณภาพของภาพ และการสกัดรูปแบบของความลึกของบุคคล
(ก) ภาพความลึก 8 บิต ที่ได้จากกล้องความลึก (ข) ส่วนที่เคลื่อนไหวที่ได้จากการตรวจจับการเคลื่อนไหว (ค) กลุ่มก้อนของวัตถุตามพิกัดของกรอบสี่เหลี่ยม (Bounding Rectangle) ที่ประกอบไปด้วยตำแหน่งบนซ้าย X_h, Y_h, ความกว้าง W_h, และความสูง H_h (ง) การตัดส่วนที่ยื่นออกมาโดยใช้เทคนิคการวิเคราะห์ภาพฉายที่เป็นไบนารี่ (จ) โครงสร้างของมนุษย์ในรูปแบบของความลึก

3.2.2 การทำแบบจำลองมนุษย์และการสกัดฟีเจอร์โดยใช้โมดูลสกัดฟีเจอร์แบบเลเยอร์จากมุมมองเดี่ยว (Layer Human Feature Extraction)

ทางผู้วิจัยได้ติดตั้งแบบจำลอง รูปแบบความลึกของมนุษย์ ลงในเลเยอร์ ที่สามารถสกัดคุณลักษณะเฉพาะของฟีเจอร์ที่ขึ้นอยู่กับ ระดับตามแนวดิ่งต่าง ๆ ของโครงสร้างมนุษย์ โดยที่คุณลักษณะทางกายภาพจะมีความแตกต่างกันในแต่ละเลเยอร์ ในแต่ละท่าทางต่าง ๆ โดยรูปแบบความลึกจะถูกตัดแบ่งในแนวดิ่งลงในจำนวนคี่ของเลเยอร์ ตัวอย่างเช่น 5 เลเยอร์ ที่แสดงไว้ในภาพประกอบที่ 3-7 โดยขนาดของเลเยอร์อันหนึ่งซึ่งจะไม่คำนึงถึงระยะทางมุมมองใด ๆ ซึ่งจะทำให้สามารถรวมฟีเจอร์ที่มีระดับเดียวกันจากทุกมุมมองเพื่อสร้างเป็นฟีเจอร์อันใหม่ที่มีความน่าเชื่อถือยิ่งขึ้น

ตัววัตถุที่อยู่ในกรอบสี่เหลี่ยมจะถูกแบ่งเท่า ๆกันลงในเลเยอร์ เพื่อที่จะแสดงฟีเจอร์ในระดับที่แตกต่างกันของโครงสร้างวัตถุ ซึ่งจะได้เป็น

$L[k]|k\in\{-N, -N+1, -N+2,..., 0, 1, 2,..., N-2, N-1, N\}$ และจำนวนทั้งหมดของเลเยอร์จะได้เป็น $2N+1$ โดยที่ N เป็นจำนวนที่สูงที่สุดของเลเยอร์ด้านบนและเลเยอร์ด้านล่าง ดังตัวอย่างในภาพประกอบที่ 3-7



(ก) (ข)

ภาพประกอบที่ 3-7 แบบจำลองของมนุษย์แบบเลเยอร์สำหรับการฟิวชันข้อมูล
(ก) ตัวอย่างในท่ายืน (ข) ตัวอย่างในท่านั่ง

โครงสร้างมนุษย์จะถูกแบ่งออกเป็น 5 เลเยอร์ (N เท่ากับ 2) โดยที่ประกอบไปด้วยเลเยอร์ด้านบน 2 เลเยอร์ เลเยอร์ด้านล่าง 2 เลเยอร์ และเลเยอร์ตรงกลาง 1 เลเยอร์ ดังนั้น เลเยอร์จะประกอบไปด้วย $\{L[-2], L[-1], L[0], L[+1], L[+2]\}$ ขอบในแนวราบจะแสดงไว้ใน

ภาพประกอบที่ 3-7 (เส้นแนวตั้ง) โดยที่ได้จากตำแหน่งด้านซ้ายและขวาของวัตถุ ส่วนขอบด้านแนวดิ่งในแต่ละเลเยอร์จะแสดงไว้ในภาพประกอบที่ 3-7 (ในเส้นแนวดิ่ง) โดยจะถูกกำหนดด้วยตำแหน่งของขอบด้านบน $y_T[\,k\,]$ และตำแหน่งของขอบด้านล่าง $y_B[\,k\,]$ ซึ่งสามารถคำนวณได้จากสมการที่ (3.19) และ (3.20)

$$y_T[k]=\frac{H_h(k+N)}{2N+1}+1 \tag{3.19}$$

$$y_B[k]=\frac{H_h(k+N+1)}{2N+1} \tag{3.20}$$

โดยพื้นที่ของเลเยอร์ที่ k ใด ๆ จะถูกกำหนดโดย $x = 0$ ถึง W_h และ $y = y_T[\,k\,]$ ถึง $y_B[\,k\,]$

ตามแบบจำลองที่ได้กล่าวมานั้น เลเยอร์ที่ได้มาจากรูปแบบของความลึกในวัตถุ จะสามารถใช้คำนวณฟีเจอร์ที่อยู่ในเลเยอร์ต่าง ๆ โดยใช้คุณสมบัติพื้นฐานและสถิติ อย่างเช่น แกนกลางความหนาแน่น ความลึก ความกว้าง พื้นที่ เป็นต้น โดยที่จะถูกเรียงตามลำดับแล้วต่อกันเป็นฟีเจอร์ ในทุก ๆ ส่วน ซึ่งความลึกสามารถช่วยในการแยกแยะคุณลักษณะเฉพาะของท่าทางต่าง ๆ ดังที่จะกล่าวจะต่อไป

ในแบบจำลองของผู้วิจัย ผู้วิจัยได้กำหนด 2 ฟีเจอร์หลักในแต่ละเลเยอร์ ประกอบไปด้วยความหนาแน่นของเลเยอร์ ($\rho[k]$) และค่าถ่วงน้ำหนักความหนาแน่นโดยความลึกของเลเยอร์ ($Z[k]$) นอกจากนี้ ค่าอัตราส่วนของวัตถุ (P_v) ซึ่งเป็นฟีเจอร์โดยรวมจะถูกเพิ่มเข้ามาเพื่อใช้เป็นตัวปรับค่า เพื่อแยกแยะจัดการกับท่าทางที่อยู่ในแนวนอน

(ก) ความหนาแน่นของเลเยอร์ ($\rho[k]$)

ความหนาแน่นของเลเยอร์ ($\rho[k]$) จะบ่งชี้จำนวนรวมของวัตถุในแต่ละเลเยอร์ ซึ่งมีความแตกต่างกันอย่างชัดเจนตามท่าทางต่าง ๆ และสามารถที่จะคำนวณได้จากพิกเซลสีขาวในแต่ละเลเยอร์ซึ่งแสดงไว้ในสมการที่ (3.21)

$$\rho'[k]=\sum_{y=y_T[k]}^{y_B[k]}\sum_{x=0}^{W_h} I_m(x,y) \tag{3.21}$$

ในมุมมองที่แตกต่างกันจะเกิดความแตกต่างกันของระยะทางระหว่างตัววัตถุและกล้องซึ่งมีผลต่อค่า $\rho'[k]$ วัตถุที่อยู่ใกล้กล้องจะปรากฏมีขนาดใหญ่กว่าวัตถุที่ไกลออกไป ดังนั้นค่าของ $\rho'[k]$ จะต้องถูกนอมัลไรซ์ เพื่อที่จะสามารถนำค่าที่มีลักษณะที่เทียบเคียงกันได้มาใช้ในการฟิว-ชันข้อมูลกันต่อไป โดยผู้วิจัยจะใช้ค่าที่มีค่ามากที่สุดในตัววัตถุ ดังที่แสดงไว้ในสมการที่ (3.22)

$$\rho[k] = \frac{\rho'[k]}{arg\,max(\rho'[k])} \tag{3.22}$$

(ข) ค่าถ่วงน้ำหนักความหนาแน่นโดยความลึกของเลเยอร์ ($z[k]$)

ค่าถ่วงน้ำหนักความหนาแน่นโดยความลึกของเลเยอร์ ($Z[k]$) จะได้มาจากการถ่วงน้ำหนักของค่าความหนาแน่นที่ถ่วงน้ำหนักโดยค่าย้อนกลับของความลึก ($Q[k]$) และค่าความหนาแน่นของเลเยอร์ ($\rho[k]$) ซึ่งค่า $Q[k]$ จะเป็นค่าที่ปรับปรุงค่าของ $\rho[k]$ ซึ่งกระบวนการหาค่า $Q[k]$ นั้นจะถูกแบ่งออกเป็นสองส่วนคือ การดึงค่าย้อนกลับของความลึก ($D_i[k]$) และการถ่วงน้ำหนักสำหรับค่าความหนาแน่นของเลเยอร์ ในขั้นตอนแรกของการดึงความลึกออกมาใช้จากเลเยอร์รูปแบบของความลึกจะเผยผิวหน้าของวัตถุที่แสดงถึงโครงสร้างในเบื้องต้น โดยมีระยะจาก 0 ถึง 255 หรือระยะทางใกล้ไปหาไกลนับจากกล้อง ตามมุมมองต่าง ๆ จะมีการแปรเปลี่ยนอยู่ในรูปแบบของโพลิโนเมี่ยม (Polynomial Form) โดยที่ค่าของความลึก ณ จุดที่ใกล้ อย่างเช่น 4-5 จะมีค่าที่แตกต่างกันน้อยตามระยะจริง ซึ่งต่างกับค่าความลึกที่อยู่ไกล เช่น จาก 250-251 จะมีค่าที่แตกต่างจากระยะจริงที่มาก ระยะห่างของความลึกจริงของเลเยอร์ ($D'[k]$) จะเป็นการกำหนดคุณสมบัติของค่าความลึกใน 2 มิติ ที่ถูกแปลงไปเป็นความลึกจริง 3 มิติ ในหน่วยของเซนติเมตร ค่าความลึกตามระยะจริงจะมีความสามารถในการแยกแยะค่าได้ดีกว่าในระหว่างเลเยอร์ของวัตถุ และยังช่วยเพิ่มความสามารถในการแยกแยะท่าทาง ผู้วิจัยได้ใช้สมการถดถอยแบบโพลิโนเมี่ยม [75] (Polynomial Regression) เพื่อแปลงค่าความลึกจริงของเลเยอร์เป็นระยะห่างตามความลึกจริง ดังสมการที่ (3.23)

$$f_c'(x) = 1.251217e^{-10}x^6 - 1.037037e^{-7}x^5 + 3.501481e^{-5}x^4 - 0.006100x^3 + 0.577595x^2 - 27.634268x + 5.675994e^2 \tag{3.23}$$

หลังจากนั้น ค่า $D'[k]$ ในแต่ละเลเยอร์จะถูกแทนโดยค่าที่ถูกแปลงของทุก ๆค่าความลึกโดยเฉลี่ยในเลเยอร์นั้น ๆ ดังสมการที่ (3.24)

$$D'[k] = f_c' \left(\frac{\sum_{y=y_T[k]}^{y_B[k]} \sum_{x=0}^{W_h} I_{mo}(x,y)}{\rho[k]} \right) \quad (3.24)$$

ค่าเกี่ยวกับตัวเลขที่สามารถนำมาเปรียบเทียบในรูปแบบความลึกของโครงสร้างมนุษย์ในจุดใด ๆของมุมมอง ซึ่งค่า ($D'[k]$) จำเป็นต้องนอมัลไรท์โดยใช้ค่าที่มากที่สุดของตัวมันเองจากทุกเลเยอร์ ก็จะได้ $D[k]$ ที่เป็นค่าความลึกจริงที่ถูกนอมัลไลท์ ดังแสดงไว้ในสมการที่ (3.25)

$$D[k] = \frac{D'[k]}{arg\,max(D'[k])} \quad (3.25)$$

ในขั้นตอนถัดไปผู้วิจัยได้ใช้ค่าย้อนกลับของความลึกระยะจริง ($D_i[k]$) (ต่อจากนี้จะเรียกว่าค่าย้อนกลับความลึก) โดยจะใช้ในการถ่วงน้ำหนักค่าความหนาแน่น ($\rho[k]$) เพื่อที่จะปรับปรุงฟีเจอร์ $\rho[k]$ สำหรับการเพิ่มความสามารถในการรู้จำของท่าทาง อย่างเช่น ท่านั่ง และท่าก้ม ค่าย้อนกลับความลึก (Inverse-depth) ดังสมการที่ (3.26) เพื่อที่จะวัดปริมาตรที่ซ่อนอยู่ภายในโครงสร้างของตัวบุคคล ซึ่งจะสามารถทำให้แยกแยะท่าทางบางท่าทางได้ดีขึ้น อย่างเช่น ดังในตารางที่ 3-3 ในการมองท่าก้มจากด้านหน้า ส่วนลำตัวด้านบนจะถูกซ่อนไว้ซึ่งค่าย้อนกลับความลึกจะสามารถเผยให้เห็นปริมาตรที่ถูกซ่อนอยู่ในช่วงนี้ออกมาได้ และในการมองท่านั่งในด้านหน้า ค่าความลึกในส่วนของต้นขาจะถูกเผยให้เห็นออกมา เมื่อเทียบกับสัดส่วนอื่น ๆ

$$D_i[k] = \left(\frac{D[k] - arg\,max(D[k])}{arg\,min(D[k]) - arg\,max(D[k])} \right) \quad (3.26)$$

ค่าความหนาแน่นที่ถ่วงน้ำหนักโดยค่าย้อนกลับของความลึก ($Q[k]$) ดังสมการที่ (3.27) ถูกกำหนดค่าโดยผลคูณของค่าย้อนกลับของความลึกระยะจริง ($D_i[k]$) และค่าความหนาแน่นของเลเยอร์ ($\rho[k]$)

$$Q[k] = (D_i[k])(\rho[k]) \quad (3.27)$$

ผู้วิจัยได้ใช้ค่าอัตราการเรียนรู้ α ที่เป็นพารามิเตอร์ที่สามารถปรับค่าได้ โดยจะทำให้สามารถถ่วงสมดุลระหว่างค่า $Q[k]$ และ $\rho[k]$ ในการปรับตั้งค่า เมื่อรูปแบบของค่าย้อนกลับของความลึกระยะจริง $D_i[k]$ เข้าใกล้ศูนย์ $Q[k]$ จะมีค่าเข้าใกล้ค่าความหนาแน่น ($\rho[k]$) ซึ่งในสมการ

ที่ (3.28) จะกำหนดค่าของ***ค่าถ่วงน้ำหนักความหนาแน่นโดยความลึกของเลเยอร์*** (Z[k]) ที่สามารถใช้ปรับค่าอัตราการเรียนรู้ α ให้ค่าใดค่ามากหรือน้อยได้

$$Z[k]=(1-\alpha)(Q[k])+(\alpha)(\rho[k]) \tag{3.28}$$

จากผลลัพธ์ของฟีเจอร์ที่ได้แสดงไว้ในตารางที่ 3-4 ***ค่าถ่วงน้ำหนักความหนาแน่นโดยความลึกของเลเยอร์*** - ซึ่งเป็นฟีเจอร์ขั้นสุดท้ายที่ใช้ในกระบวนการรู้จำท่าทาง สามารถปรับปรุงรูปแบบของฟีเจอร์โดยทำให้สามารถแยกแยะได้ดีขึ้น ถึง 13 รูปแบบ จาก 20 รูปแบบของฟีเจอร์ และไม่เกิดความเปลี่ยนแปลง 5 รูปแบบฟีเจอร์ (I ถึง V ในท่าทางยืน – เดิน) และมี 2 รูปแบบของฟีเจอร์ที่แย่ลง (VIII ในท่านั่งจากมุมมองด้านข้าง และ X ในท่านั่งมุมมองจากด้านหลัง) ซึ่งโดยรวมแล้ว Z[k] ทำให้รูปแบบของฟีเจอร์ที่ดีขึ้น แม้ว่าจะมีบางรูปแบบของ Z[k] ไม่เปลี่ยนแปลง และมีบางรูปแบบที่แย่ลงเล็กน้อยเป็นส่วนน้อย ซึ่งต้องอาศัยกระบวนการในขั้นถัดไปที่เป็นการฟิวชันข้อมูลจากหลายมุมมองเพื่อที่จะทำให้รูปแบบเหล่านี้ สามารถแยกแยะออกจากกันได้

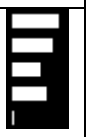
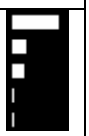
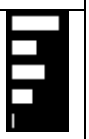
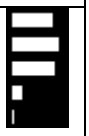
(ค) ค่าอัตราส่วนของวัตถุ (P_v)

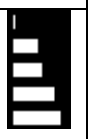
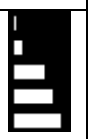
ค่าอัตราส่วนของวัตถุ (P_v) เป็นพารามิเตอร์ Penalty ของแบบจำลองที่ใช้ในการบ่งบอกสัดส่วนในเบื้องต้นของวัตถุ ซึ่งจะใช้ในการแยกแยะท่าทางที่อยู่ในแนวนอนออกจากท่าทางที่อยู่ในแนวตั้ง โดยค่าของมันจะเป็นสัดส่วนของความกว้างและความสูง ขึ้นอยู่กับตัววัตถุในแต่ละมุมมอง ดังสมการที่ (3.29)

$$P_V=\frac{W_h}{H_h} \tag{3.29}$$

ในตารางที่ 3-3 ที่ได้กล่าวถึงก่อนหน้านี้ได้แสดงให้เห็นถึงรูปแบบความลึกในแต่ละท่าทางและฟีเจอร์ของตัวมันเองในแตะละมุมมอง โดยทั่วไปแล้วรูปแบบของฟีเจอร์ในแต่ละท่าทางจะมีค่าเหมือนกันในแต่ละท่าทางนั้น ๆ แม้ว่าจะมีบางส่วนที่แตกต่างกันไปบ้างเล็กน้อย โดยขึ้นอยู่กับมุมมอง ซึ่งรูปแบบที่เกิดขึ้นได้จากกล้องหลายตัว ซึ่งมีตำแหน่ง ณ ความสูง 2 เมตรจากพื้น โดยชี้ลงไปทำมุม 30 องศากับแนวระดับ

ตารางที่ 3-3 ตัวอย่างรูปแบบของความลึกในตัวบุคคลรวมไปถึงฟีเจอร์ในแต่ละมุมมองและท่าทาง

ท่าทาง / มุมมอง	I_{mo}	ρ [k]	D [k]	D_i[k]	ท่าทาง / มุมมอง	I_{mo}	ρ [k]	D [k]	D_i[k]
ยืน-เดิน / หน้า					ก้ม / หน้า				
ยืน-เดิน / เฉียงทางหน้า					ก้ม / เฉียงทางหน้า				
ยืน-เดิน / ข้าง					ก้ม / ข้าง				
ยืน-เดิน / เฉียงทางหลัง					ก้ม / เฉียงทางหลัง				
ยืน-เดิน / หลัง					ก้ม / หลัง				
นั่ง / หน้า					นอน / หน้า				
นั่ง / เฉียงทางหน้า					นอน / เฉียงทางหน้า				
นั่ง / ข้าง					นอน / ข้าง				

ท่าทาง / มุมมอง	I_{mo}	ρ [k]	D [k]	D_i[k]	ท่าทาง / มุมมอง	I_{mo}	ρ [k]	D [k]	D_i[k]
นั่ง / เฉียงทางหลัง					นอน / เฉียงทางหลัง				
นั่ง / หลัง					นอน / หลัง				
โดยที่ I) ρ[k] คือ ค่าความหนาแน่นของเลเยอร์ II) D[k] คือ ค่าความลึกจริงที่ถูกนอมัลไลท์ III) D_i[k] คือ ค่าย้อนกลับของความลึก									

ท่าทางทั้งหมดในตารางที่ 3-3 ซึ่งได้แก่ ท่ายืน/เดิน ท่านั่ง ท่าก้ม และท่านอน จะแสดงรายละเอียดต่าง ๆ ดังนี้ สำหรับค่ายืนและเดินจะมีฟีเจอร์ที่ไม่เปลี่ยนแปลงทั้งความหนาแน่น ρ[k] และค่าย้อนกลับความลึก (D′ [k]) ส่วนค่าความลึกจริงที่ถูกนอมัลไลท์ (D[k]) จะมีความลาดชันที่เท่า ๆกันในทุก ๆมุมมอง เนื่องจากตำแหน่งของกล้อง สำหรับท่านั่งฟีเจอร์จะแปรเปลี่ยนในหลายรูปแบบของ ρ[k] และ D[k] ตามมุมมองต่าง ๆ อย่างไรก็ตามรูปแบบฟีเจอร์ของท่านั่งในมุมด้านหน้าและมุมเฉียงจะมีค่าค่อนข้างคล้ายคลึงกับท่าทางในท่ายืน/เดิน โดยที่ค่า D[k] ในบางมุมมองจะบ่งบอกปริมาตรที่ซ่อนอยู่ในส่วนของต้นขา สำหรับท่าก้มจะมีรูปแบบของ ρ[k] ที่ค่อนข้างเหมือนกันในแต่ละมุมมอง ยกเว้นในด้านหน้าและหลัง เนื่องจากการบดบังในส่วนของลำตัวในท่อนบน อย่างไรก็ตามค่า D_i[k] สามารถแสดงแทนถึงพื้นที่ที่ถูกบดบังได้อย่างชัดเจน และสำหรับท่านอนจะมีค่า ρ[k] ที่แปรปรวนในแต่ละมุมมองซึ่งยากที่จะแยกแยะโดยการใช้พื้นฐานของฟีเจอร์เลเยอร์ ซึ่งในกรณีเฉพาะนี้จึงใช้ค่าอัตราส่วนของวัตถุ (P_r) ในที่เพิ่มเข้ามาช่วยในการรู้จำท่าทางในแนวนอน

3.2.2 การฟิวชันกันของฟีเจอร์แบบเลเยอร์ (Layer Feature Fusion)

ในส่วนของการฟิวชันกันของฟีเจอร์แบบเลเยอร์จะมุ่งเน้นไปในส่วนของการนำฟีเจอร์จากหลายมุมมองที่ต่างกันมารวมกัน เนื่องจากรูปแบบของฟีเจอร์สามารถที่จะเปลี่ยนแปลงหรือบดบังโดยอวัยวะของตัวบุคคลเองไปตามมุมมองต่าง ๆ ซึ่งการรู้จำจากภาพมุมมองเดี่ยวอาจจะเกิดปัญหาเหล่านี้ ซึ่งนำไปสู่การได้มาซึ่งฟีเจอร์ที่มีความคล้ายคลึงกันหรือไม่ชัดเจนในแต่ละท่าทาง

ดังนั้นผู้วิจัยจึงได้นำเสนอวิธีการที่จะฟิวชันฟีเจอร์จากหลายมุมมองเข้าด้วยกันเพื่อที่จะปรับปรุงการให้ดียิ่งขึ้น โดยแต่ละมุมมองเดี่ยวจะมี 3 ฟีเจอร์ที่ถูกดึงออกมาเพื่อใช้ในการฟิวชันข้อมูล ได้แก่ ความหนาแน่นของเลเยอร์ ($\rho[k]$), ค่าถ่วงน้ำหนักความหนาแน่นโดยความลึกของเลเยอร์ ($Z[k]$), และค่าอัตราส่วนของวัตถุ (P_v) ผู้วิจัยได้สร้างสองฟีเจอร์ฟิวชันโดยเป็นการฟิวชันกันเพื่อระบุคุณลักษณะใหม่ในเชิงของปริมาตรและพื้นที่ของโครงสร้างของร่างกายจากทุก ๆมุมมองในแต่ละเลเยอร์ ได้แก่ ***ค่ามวลสารของมิติ-Mass of Dimension*** (ω[k]) และ***ค่ามวลสารของมิติแบบถ่วงน้ำหนักจำเพาะ-Weighted Mass of Dimension*** ($\overline{\omega}$[k])

(ก) ค่ามวลสารของมิติ (Mass of Dimension)

ค่ามวลสารของมิติ ($\omega[k]$) จะนำมาจากผลคูณรวมของค่าความหนาแน่นของเลเยอร์ ($\rho[k]$) จากทุกมุมมองที่มี จาก $v = 1$ (มุมมองจากกล้องแรก) จนถึง $v = d$ (มุมมองจากกล้องสุดท้าย/จำนวนของมุมมอง) ตามสมการที่ (3.30)

$$\omega[k] = \prod_{v=1}^{d} (\rho_{(v=i)}[k]) \tag{3.30}$$

(ข) ค่ามวลสารของมิติแบบถ่วงน้ำหนักจำเพาะ (Weighted Mass of Dimension)

ค่ามวลสารของมิติแบบถ่วงน้ำหนักจำเพาะ ($\overline{\omega}[k]$) ถูกกำหนดขึ้นตามสมการที่ (3.31) โดยที่เป็นผลคูณรวมของค่าถ่วงน้ำหนักความหนาแน่นโดยความลึกของเลเยอร์ ($Z[k]$) จากทุก ๆ มุมมองที่มี

$$\bar{\omega}[k] = \prod_{v=1}^{d} \left(Z_{(v=i)}[k]\right) \tag{3.31}$$

นอกจากนี้ ค่าสูงสุดของค่าอัตราส่วนของวัตถุที่ถูกเลือกจากทุกมุมมอง (สมการที่ (3.32)) ซึ่งจะถูกใช้ในการสร้างเวกเตอร์ของฟีเจอร์ในส่วนถัดไป

$$P_m = \text{argmax}\ (P_{v(v=1)}, P_{v(v=2),} P_{v(v=3),\ldots,} P_{v(v=d)}) \tag{3.32}$$

ถัดจากนั้นจะเป็นส่วนของการสร้างเวกเตอร์ของฟีเจอร์ (Feature Vector) โดยจะแบ่งเป็น 2 ประเภท ตามการนำไปใช้งาน ได้แก่ แบบไม่ใช้ความลึก (Non-depth Considering) และแบบใช้ความลึก (Depth Considering) ซึ่งเวกเตอร์ของฟีเจอร์เหล่านี้จะถูกนำไปใช้ในขั้นตอนการเรียนรู้จดจำและแยกแยะในขั้นตอนถัดไปเพื่อสามารถจะนำมาใช้ในการรู้จำท่าทางต่อไป

(ค) เวกเตอร์ของฟีเจอร์แบบไม่ใช้ความลึก (Non-depth Considering Feature Vector)

เวกเตอร์ของฟีเจอร์แบบที่ไม่ใช้ความลึกจะสร้างขึ้นมาจากการเรียงต่อกันของค่ามวลสารของมิติ (Mass of Dimension) จากทุกเลเยอร์ ซึ่งได้มาจากผลคูณรวมของค่าความหนาแน่นของเลเยอร์ ($\rho[k]$) และนำมาต่อกับค่ามากที่สุดจากทุกมุมมองของค่าอัตราส่วนของวัตถุ (P_v) ซึ่งจะได้เป็นค่าดังนี้ $P_m = \text{argmax}\,(P_{v(v=i)})$ เมื่อนำมาเรียงต่อกันจะได้ข้อมูลที่เรียงต่อกันตามตัวอย่างต่อไปนี้

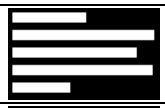
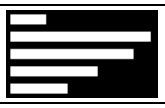
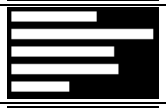
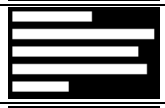
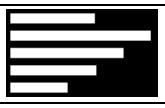
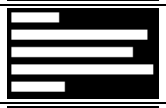
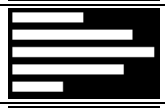
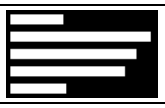
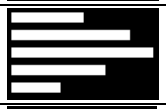
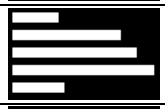
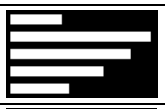
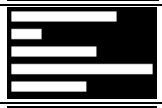
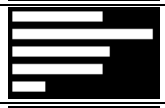
$$\left\{\begin{matrix}\omega[-N],\ \omega[-N+1],\ \omega[-N+2],\ \ldots,\ \omega[0],\ \omega[1],\ \omega[2],\\ \ldots,\ \omega[N],\ P_m\end{matrix}\right\}$$

(ง) เวกเตอร์ของฟีเจอร์แบบความลึก (Depth Considering Feature Vector)

เวกเตอร์ของฟีเจอร์แบบความลึกจะได้จากการเรียงต่อกันของค่ามวลสารของมิติแบบถ่วงน้ำหนักจำเพาะ ($\overline{\omega}[k]$) ประกอบกับค่ามากที่สุดจากทุกมุมมองของค่าอัตราส่วนของวัตถุ (P_m) ซึ่งจะได้เวกเตอร์ของฟีเจอร์แบบความลึกดังต่อไปนี้

$$\left\{\begin{matrix}\bar{\omega}[-N],\ \bar{\omega}[-N+1],\ \bar{\omega}[-N+2],\ \ldots,\ \bar{\omega}[0],\ \bar{\omega}[1],\ \bar{\omega}[2],\\ \ldots,\ \bar{\omega}[N],\ P_m\end{matrix}\right\}$$

ตารางที่ 3-4 ตัวอย่างฟีเจอร์ในมุมมองเดี่ยวและฟีเจอร์ฟิวชันที่ได้จากหลายมุมมองจากกล้องหลายตัว

ท่าทาง / มุมมอง	กล้องที่ 1		กล้องที่ 2		$\overline{\omega}[k]$
	$\boldsymbol{I}_{mo}$	$Z[k]\|\alpha=0.9$	$\boldsymbol{I}_{mo}$	$Z[k]\|\alpha=0.9$	
I. ยืน-เดิน / หน้า					
II. ยืน-เดิน / เฉียงทางหน้า					
III. ยืน-เดิน / ข้าง					
IV. ยืน-เดิน / เฉียงทางหลัง					
V. ยืน-เดิน / หลัง					
VI. นั่ง / หน้า					
VII. นั่ง / เฉียงทางหน้า					
VIII. นั่ง / ข้าง					
IX. นั่ง / เฉียงทางหลัง					
X. นั่ง / หลัง					
XI. ก้ม / หน้า					
XII. ก้ม / เฉียงทางหน้า					
XIII. ก้ม / ข้าง					
XIV. ก้ม / เฉียงทางหลัง					
XV. ก้ม / หลัง					

ท่าทาง / มุมมอง	กล้องที่ 1		กล้องที่ 2		$\overline{\omega}[k]$
	$\boldsymbol{I}_{mo}$	$Z[k]\|\alpha=0.9$	$\boldsymbol{I}_{mo}$	$Z[k]\|\alpha=0.9$	
XVI. นอน / หน้า					
XVII. นอน / เฉียงทางหน้า					
XVIII. นอน / ข้าง					
XIX. นอน / เฉียงทางหลัง					
XX. นอน / หลัง					

โดยที่: I) $Z[k]$ คือ ค่าถ่วงน้ำหนักความหนาแน่นโดยความลึกของเลเยอร์ ($Z[k]$) โดยในตัวอย่างตั้งค่าอัตราการเรียนรู้ α ซึ่งปรับค่าระหว่าง $Q[k]$ และ $\rho[k]$ ให้เท่ากับ 0.9 II) $\overline{\omega}[k]$ คือ ค่ามวลสารของมิติแบบถ่วงน้ำหนักจำเพาะ ($\overline{\omega}[k]$)

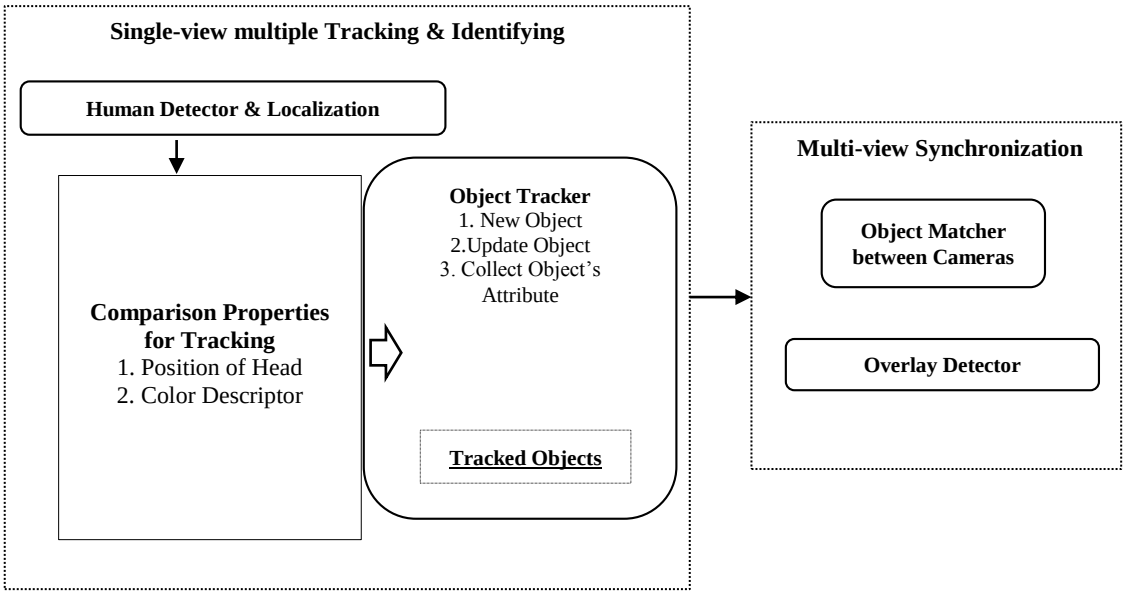
จากตารางที่ 3-4 ได้แสดงค่าของมวลสารของมิติแบบถ่วงน้ำหนักจำเพาะ ($\overline{\omega}[k]$) ที่ฟิวชันข้อมูลจากค่าถ่วงน้ำหนักความหนาแน่นโดยความลึกของเลเยอร์ ($Z[k]$) โดยมีสองมุมมอง (ใช้กล้อง 2 ตัวในการทดลอง) โดยรูปแบบข้อมูลที่ถูกฟิวชันกัน (Fused Feature) ในท่ายืน-เดินในแต่ละมุมมองมีลักษณะที่ใกล้เคียงกันมาก สำหรับท่านั่งโดยทั่วไปแล้วจะมีรูปแบบที่ไปในทิศทางเดียวกันแม้จะมีความแตกต่างกันอยู่บ้าง ยกเว้นในท่านั่งจากด้านหลังจะแตกต่างไปเนื่องจากข้อมูลในส่วนล่างหายไปเนื่องจากถูกบดบัง ซึ่งรูปแบบข้อมูลเหล่านี้ยังคงสามารถใช้ในการรู้จำได้โดยส่วนของต้นขาซึ่งมีค่าที่แสดงปริมาตรออกมามากกว่าส่วนอื่น ในส่วนของท่าก้มนั้น ตัวรูปแบบของข้อมูลที่ถูกฟิวชันกันมีความสอดคล้องกันมากในทุก ๆมุมมอง โดยมีลักษณะที่ไปกองรวมอยู่ที่ช่วงเลเยอร์ด้านบนแต่จะมีความแตกต่างกันบ้างในส่วนของขนาดของความโค้ง และในส่วนของท่านอนจะมีความแตกต่างกันอยู่มากโดยชัดเจนในมุมมองต่าง ๆโดยเฉพาะในส่วนของมุมมองด้านหน้า ซึ่งอย่างไรก็ตามก็ยังพอจะมีคุณลักษณะในส่วนของเลเยอร์ด้านบนที่มีค่ามากและลดหลั่นลงไปตามลำดับของเลเยอร์ ซึ่งในท่านั่งก็ยังมีค่ามากที่สุดจากทุกมุมมองของค่าอัตราส่วนของวัตถุ (P_{m}) ซึ่งจะมีความแตกต่างชัดเจนกับท่าอื่น ๆที่มาช่วยใช้ในการรู้จำท่าทาง

3.3 การติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง (Multi-view Human Tracking and Person Re-identification)

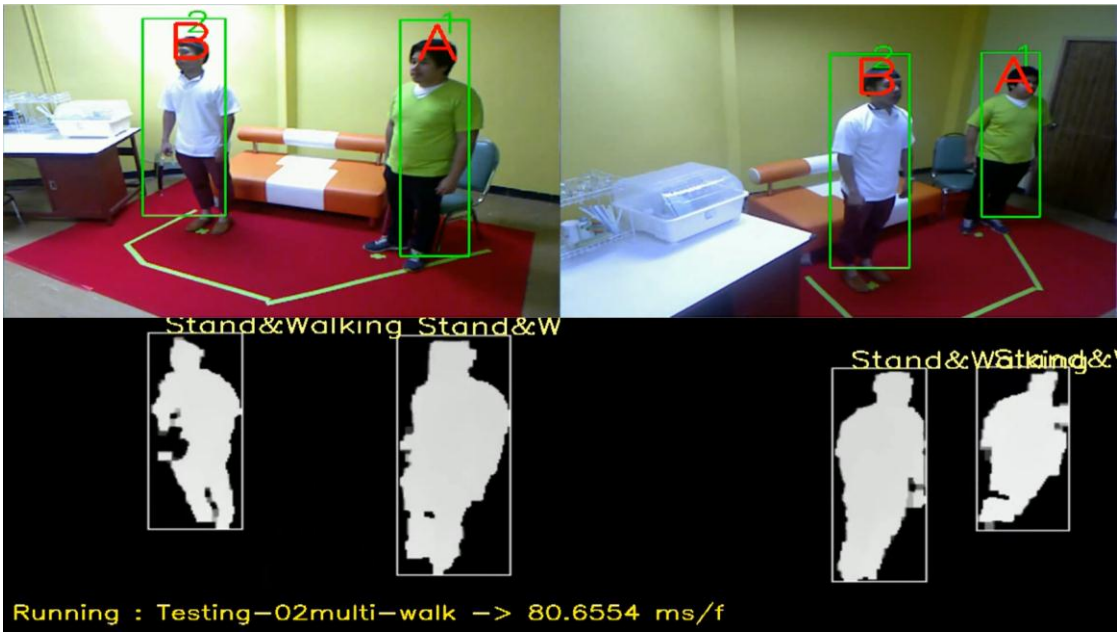
การติดตามและจดจำตัวบุคคลเป็นสิ่งที่จำเป็นที่ต้องมีในระบบเฝ้าระวังอัจฉริยะต่าง ๆที่ทำงานเกี่ยวกับการวิเคราะห์กับมนุษย์ เพื่อที่จะติดตามเมื่อเข้ามาในบริเวณที่ระบบทำการวิเคราะห์ทั้งตำแหน่งและสถานะต่าง ๆในแต่ละเฟรมที่ทำการวิเคราะห์ (Tracking), การติดตามและจับคู่บุคคลเดียวกันกรณีที่มีบุคคลเข้ามามากกว่าหนึ่งคนในซึ่งใช้หลายกล้องเพื่อวิเคราะห์ (Matching) และเพื่อระบุในเบื้องต้นว่ามีบุคคลใดบ้างที่เข้า-ออกในบริเวณที่กำลังวิเคราะห์อยู่ (Re-identification)

สำหรับในระบบการติดตามและจดจำตัวบุคคลจะใช้ข้อมูลตำแหน่งและสี ซึ่งสีเป็นข้อมูลเบื้องต้นที่เด่นชัดที่สามารถนำมาใช้ในการแยกแยะแต่ละบุคคลทั้งในกล้องเดียวกันและระหว่างกล้อง รวมถึงใช้ในการจดจำตัวบุคคลในขั้นต้น แทนการใช้ใบหน้า โครงร่าง และลักษณะทางกายภาพอื่น ๆของมนุษย์ซึ่งต้องใช้วิธีการบางอย่าง เช่น เดินมายังจุดที่มองเห็นใบหน้าได้อย่างชัดเจนในมุมมองตรงเพื่อที่จะรู้จำเพื่อจดจำตัวบุคคล และปัจจัยด้านเวลาและความซับซ้อนในการประมวลผลซึ่งบางครั้งไม่สามารถทำได้ในการประมวลแบบ Real-time

ในระบบจะประกอบไปด้วย 2 ส่วน คือ ส่วนติดตามบุคคลในแต่ละมุมมอง (Single-view Multiple Tracking and Identifying Module) โดยทำหน้าที่ติดตามและจดจำตัวบุคคลจากกล้องเดี่ยวในแต่ละกล้อง และส่วนการเชื่อมโยงข้อมูลระหว่างมุมมองกล้อง (Multi-view Synchronizing Module) ที่ทำหน้าที่เพื่อจับคู่บุคคลระหว่างกล้องและดึงข้อมูลที่ต้องใช้ออกมา รวมไปถึงการตรวจจับการเกิดการบดบัง ดังภาพประกอบที่ 3-8



ภาพประกอบที่ 3-8 ภาพรวมระบบการติดตามและจดจำตัวบุคคล



ภาพประกอบที่ 3-9 ตัวอย่างผลลัพธ์การทำงานของระบบการติดตามและจดจำตัวบุคคล

จากภาพประกอบที่ 3-9 แสดงให้เห็นถึงการทำงานของระบบโดยแบ่งออกเป็นส่วนติดตามแต่ละบุคคลในกล้องเดี่ยว โดยจะใช้เลขในการติดตามและจดจำตัวบุคคลในกล้องเดี่ยวโดยเรียงตามลำดับการเข้าก่อน-หลัง ซึ่งกรอบสีเขียวหมายถึงสถานะ Active คือเข้ามาอยู่ในตัวกล้องและสามารถติดตามและจดจำตัวบุคคลได้ ส่วนศูนย์รวมเชื่อมโยงข้อมูลจะทำหน้าที่ในการ Assign ข้อมูลในการจดจำให้กับตัวบุคคล โดยใช้ ID A-Z และจะทำการ Matching บุคคลระหว่างกล้อง

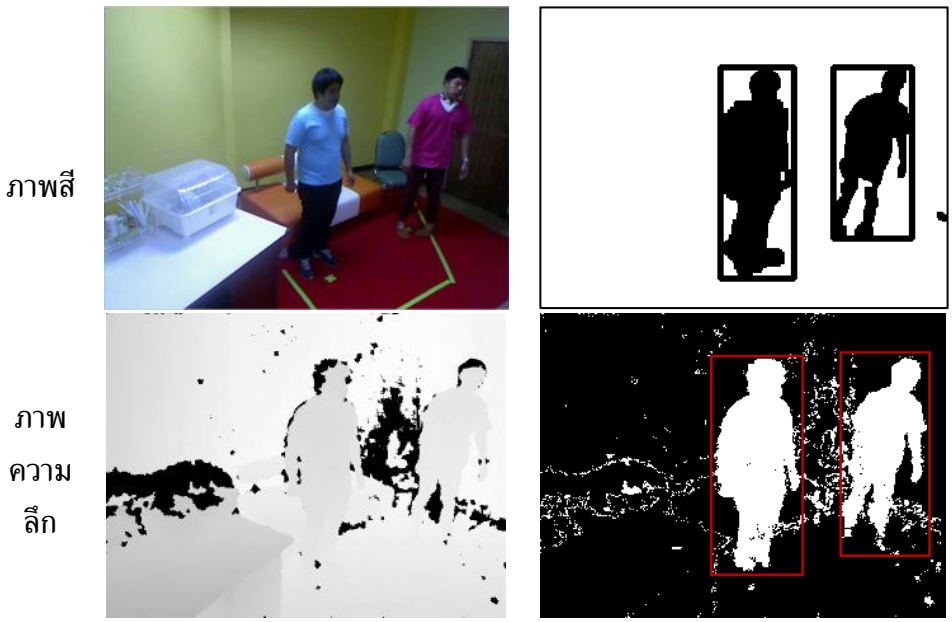
ดังตัวอย่างในภาพประกอบจะ Assign ID เป็น A และ B (สีแดง) เพื่อใช้ในการติดตามและจดจำตัวบุคคลทั้งสอง

3.3.1 ส่วนการติดตามบุคคลในแต่ละมุมมอง (Single-view Tracking and Identifying Module)

จากภาพประกอบที่ 3-8 ส่วนการติดตามบุคคลในแต่ละมุมมอง (Single-view Tracking and Identifying Module) ประกอบไปด้วย ตัวตรวจจับบุคคลในกล้อง (Human Detector) ส่วนต่อมาจะเป็นส่วนที่รับข้อมูลจากตัวตรวจจับบุคคลเพื่อนำมาดึงคุณลักษณะที่ใช้ในการเปรียบเทียบกับข้อมูลของบุคคลที่มีอยู่ในระบบเพื่อติดตามและระบุตัวตน ซึ่งในส่วนการติดตามบุคคล (Object Tracker) จะนำข้อมูลที่ได้มาเปรียบเทียบทั้ง Color Descriptor และ Position of Head เพื่อใช้ในการเปรียบเทียบและอัพเดทกับบุคคลเดิมที่ติดตามอยู่ หากเปรียบเทียบแล้วคุณลักษณะต่างกันและมีแนวโน้มว่าไม่รู้จักบุคคลที่ได้จากตรวจจับบุคคลที่เข้ามาใหม่ก็จะ Assign เป็นบุคคลใหม่ในระบบเพื่อติดตามและจดจำ และยังมีการ Destroy กลุ่มข้อมูลที่เป็น Noise รวมไปถึงการนำข้อมูลคุณสมบัติต่าง ๆของบุคคลที่ใช้ในรู้จำท่าทางและอื่น ๆ เข้าไปเก็บในข้อมูลของบุคคลที่ระบบกำลังติดตามอยู่

(ก) การตรวจจับตัวบุคคลและระบุตำแหน่งของบุคคลทั้งในโหมดสีและความลึก (Human Detection and Localization)

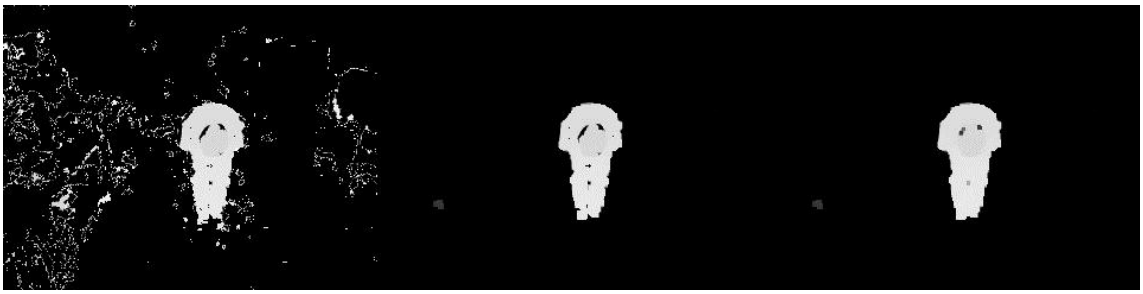
สำหรับการตรวจจับบุคคลจะใช้การตรวจจับการเคลื่อนไหว โดยการสร้างแบบจำลองสำหรับภาพพื้นหลังของทั้งภาพความลึกและภาพสี เพื่อเปรียบเทียบกับภาพที่ต้องการตรวจหาความเคลื่อนไหว และตั้งสมมุติฐานว่าวัตถุที่เคลื่อนไหวในเฟรมวิดีโอเป็นมนุษย์ทั้งหมด สำหรับการตรวจจับการเคลื่อนไหวในงานวิจัยนี้ใช้การตรวจจับการเคลื่อนไหวที่สามารถสร้างพื้นหลังจากวิธีเกาส์เซียนหลายรูปแบบ เพื่อจะเรียนรู้แบบจำลองของพื้นหลังที่ซับซ้อนได้ดีมากยิ่งขึ้น ซึ่งตัวอย่างของการตรวจจับบุคคลได้แสดงไว้ดังภาพประกอบที่ 3-10



(ก) เฟรมที่ต้องการตรวจจับบุคคล (ข) ผลลัพธ์การตรวจจับบุคคล

ภาพประกอบที่ 3-10 ตัวอย่างการตรวจจับบุคคล

ส่วนถัดมาคือการลดทอนสัญญาณรบกวน (Noise Reduction) ซึ่งการลดทอนสัญญาณรบกวนจะใช้กระบวนการทางสัณฐานวิทยาที่เป็นการประมวลผลภาพโดยการเปลี่ยนแปลงลักษณะรูปร่างหรือโครงสร้างของภาพ โดยในการปรับปรุงภาพของระบบนี้จะเน้นไปที่การใช้ Opening และ Closing ซึ่งจะไปกอบไปด้วย Dilation และ Erosion ดังภาพประกอบที่ 3-11



Input (Original Image) After Opening After Closing

ภาพประกอบที่ 3-11 ตัวอย่างของการ Closing และ Opening

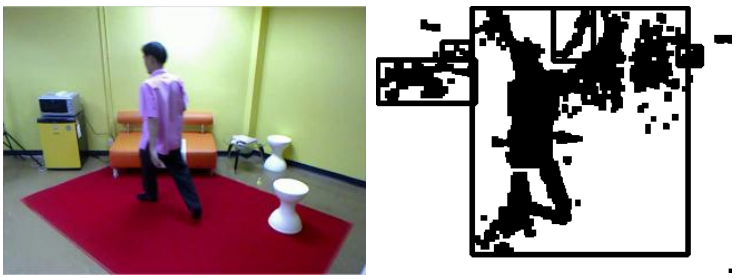
เมื่อได้โครงร่างของตัวบุคคลก็จะทำการระบุตำแหน่งของบุคคล (Object Localization) สำหรับการระบุตำแหน่งของบุคคล (Motion-Depth Object) จะใช้กระบวนการหาขอบของวัตถุ (Edge Detection) และหาสิ่งที่อยู่รอบวัตถุที่ผู้วิจัยสนใจที่เรียกว่า Contour คือ จุดที่ล้อมรอบตัววัตถุ หลังจากนั้นจะทำการประมาณจุดที่อยู่รอบวัตถุให้เหลือเฉพาะจุดที่สำคัญ (Control Point)

แล้วทำการหา Bounding Box of Object หรือที่เรียกว่ากรอบสี่เหลี่ยมของวัตถุ ในขั้นตอนแรกจะเริ่มต้นที่การหาขอบของวัตถุโดยใช้ Laplacian Edge Detection โดยวิธีนี้จะทำให้ความคมของขอบมีมากขึ้น จะใช้สำหรับทำให้ขอบของภาพมีความชัดเจนขึ้น โดยใช้ Second Derivative ซึ่งจะเห็นความแตกต่างชัดเจนกว่า First Derivative และใช้ Gaussian Smoothing Filter เพื่อกรองสัญญาณรบกวนออก เพื่อให้ได้ขอบจริงของภาพ หลังจากได้ขอบของวัตถุแล้วจะต้องทำการหาตำแหน่งของจุดที่ประกอบรวมแล้วได้เป็นขอบของวัตถุ ซึ่งเรียกว่า Contour โดยการใช้ Freeman chains เพื่อหาจุดที่ประกอบกันแล้วได้เป็นตัววัตถุ

เมื่อได้ Contour ของวัตถุแล้วจะต้องประมาณจุดที่สำคัญ (Control Point) โดยใช้ Polygon Approximations ของ Douglas-Peucker [93] โดยใช้วิธีการลากเส้นจากจุดสองจุดแล้วหาระยะห่างจากเส้นไปยังจุดต่าง ๆแล้วแบ่ง Segment ใหม่โดยใช้จุดที่มีค่ามากที่สุดจนระยะห่างน้อยกว่าค่า Threshold จึงหยุด จุดที่ได้จากกระบวนการนี้จะเป็นจุดที่สำคัญ (Control Point) แล้วจึงทำการหา Bounding Box of Object โดยคำนวณจุดซ้ายบนจากค่าน้อยที่สุด และจุดขวาล่างจากค่ามากที่สุด แล้วจึงหาค่า X, Y, Width, และ Height ซึ่งเป็น Bounding Box ของวัตถุ ดังผลลัพธ์ตามภาพประกอบที่ 3-12

ภาพประกอบที่ 3-12 ตัวอย่าง Bounding Box ของวัตถุ(บุคคล)

เนื่องจากการหาตรวจจับตัวบุคคลในโหมดภาพสี RGB จะเกิดสัญญาณรบกวนในกรณีที่แสงเปลี่ยนฉับพลัน ตามภาพประกอบที่ 3-13 จึงทำให้ตำแหน่งของตัวบุคคลที่ได้จากโหมดภาพสี RGB เชื่อถือไม่ได้ จึงต้องใช้ข้อมูลตำแหน่งจากโหมดความลึกประกอบในการหาตำแหน่งตัวบุคคลในภาพ RGB ที่ใช้ในการดึงข้อมูลสีเพื่อติดตามและจดจำตัวบุคคล



ภาพประกอบที่ 3-13 ตัวอย่างสัญญาณรบกวนเนื่องมาจากแสงทำให้ Motion Detection เกิด Noise

ในตำแหน่งของบุคคลทั้งภาพสีและความลึกจะมีขนาดและตำแหน่งที่ต่างกัน แสดงตัวอย่างตามภาพประกอบที่ 3-14 เนื่องมาจากตำแหน่งของเซ็นเซอร์ภาพสีและภาพความลึกต่างกัน จึงต้องทำการเทียบเคียงตำแหน่งด้วยวิธีการชดเชยด้วยอัตราส่วนที่คงที่ และหากตำแหน่งของความลึกห่างจากจุดศูนย์กลางจอก็ยิ่งทำให้ตำแหน่งคาดเคลื่อนจากตำแหน่งของบุคคลในภาพสีเพิ่มมากขึ้น จึงต้องมีการสร้างทั้งฟังก์ชันความเปลี่ยนแปลงในแนวราบแบบคงที่, ค่าความเปลี่ยนแปลงในแนวราบตามสัดส่วนของพิกัดแนวราบ และค่าความเปลี่ยนแปลงในความกว้างแบบคงที่ ซึ่งในงานวิจัยนี้จะไม่ใช้การปรับค่าตามความสัมพันธ์ในเชิง Geometric Calibration ที่ใช้ค่าพารามิเตอร์ของกล้องต่าง ๆ เพื่อหา Translation Matrix และ Rotation Matrix เพื่อ Matching ตำแหน่งจากภาพสีและความลึก เนื่องจากค่าพารามิเตอร์ของกล้องต่าง ๆมีค่าไม่เท่ากัน ต้องมีการทำการสอบเทียบค่าเพื่อหาค่าต่าง ๆ และใช้เวลาในการประมวลผลที่มาก และงานชิ้นนี้ต้องการเพียงตำแหน่งเบื้องต้น จากการ Shift ค่าโดยตรง และใช้ข้อมูลจากตำแหน่งของบุคคลจากภาพสีที่มีความน่าเชื่อถือมาใช้ประกอบกัน ซึ่งทำได้เร็วและค่อนข้างแม่นยำ

ภาพประกอบที่ 3-14 ขนาดและตำแหน่งที่ต่างกันของภาพสีและความลึก

ในขั้นตอนนี้จะใช้ข้อมูลของวัตถุที่เป็นตำแหน่งของบุคคลจากความลึกและ Shift ค่าไปยังตำแหน่งของบุคคลในภาพสี โดยลำดับแรกจะใช้ค่าความเปลี่ยนแปลงในแนวราบแบบคงที่ - Fixed Horizontal Shift Value ($\hbar$) สามารถคำนวณได้จากการคูณกันของความกว้างของภาพ (col) และอัตราส่วนเปลี่ยนแปลงในแนวดิ่ง - Horizontal Shift Factor (α) ซึ่งเป็นค่าคงที่ที่ได้จากการทดลองมีค่าที่เหมาะสมที่ $\alpha = 5$ แสดงตามสมการที่ (3.33)

$$\hbar = \left(col \times \left(\frac{\alpha}{100} \right) \right) \tag{3.33}$$

ค่าความเปลี่ยนแปลงในแนวราบตามสัดส่วนของพิกัดแนวราบ - Position-based Horizontal Shift Value ($\mathcal{T}$) สามารถคำนวณหาได้จากสัดส่วนการเบี่ยงเบนจากกลางภาพ ($\frac{col}{2}$) เมื่อเทียบกับตำแหน่งในแนวราบเดิม ($\mathcal{X}_{old}$) คูณด้วยค่าเปอร์เซ็นต์ตัวคูณการเปลี่ยนแปลงตามสัดส่วนของพิกัดแนวราบ - Percentage of Position-based Horizontal Shift (P) ซึ่งเป็นค่าคงที่ที่ได้จากการทดลองมีค่าที่เหมาะสมที่ $P = 10$ แสดงตามสมการที่ (3.34)

$$\mathcal{T} = -\left(\frac{\mathcal{X}_{old} - \frac{col}{2}}{\frac{col}{2}} \times P \right) \tag{3.34}$$

ค่าความเปลี่ยนแปลงในความกว้างแบบคงที่ - Width Shift Value (φ) สามารถคำนวณได้จากการคูณกันของความกว้างของภาพ (col) และอัตราส่วนเปลี่ยนแปลงในความกว้าง - Width Shift Factor (β) ซึ่งเป็นค่าคงที่ที่ได้จากการทดลองมีค่าที่เหมาะสมที่ $\varphi = 30$ แสดงตามสมการที่ (3.35)

$$\varphi = col \times \left(\frac{\beta}{100} \right) \tag{3.35}$$

จากนั้นหาค่าตำแหน่งในแนวราบใหม่ ($\mathcal{X}_{new}$) ซึ่งได้จากการรวมกันของตำแหน่งในแนวราบเดิม ($\mathcal{X}_{old}$) ค่าความเปลี่ยนแปลงในแนวดิ่งแบบคงที่ - Fixed Horizontal Shift Value ($\hbar$) และค่าความเปลี่ยนแปลงในแนวดิ่งตามสัดส่วนของพิกัดแนวราบ - Position-based Horizontal Shift Value ($\mathcal{T}$) แสดงตามสมการที่ (3.36)

$$\mathcal{X}_{new} = \mathcal{X}_{old} + \hbar + \mathcal{T} \tag{3.36}$$

ต่อมาหาค่าขนาดความกว้างของตัวบุคคลใหม่ ($\mathcal{w}_{new}$) ที่ได้จากการรวมกันของความกว้างเดิม ($\mathcal{w}_{old}$) และค่าความเปลี่ยนแปลงในความกว้างแบบคงที่ - Width Shift Value (φ) แสดงตามสมการที่ (3.37)

$$\mathcal{w}_{new} = \mathcal{w}_{old} + \varphi \tag{3.37}$$

เมื่อได้ตำแหน่งใหม่ของบุคคลในภาพก็จะเข้าสู่กระบวนการวัดค่าความน่าเชื่อถือของข้อมูลตำแหน่งของบุคคลจากภาพสี เพื่อนำไปพิจารณาใช้ประกอบในการปรับปรุงตำแหน่งที่ได้จากการ Shift ค่าโดยตรง แต่เนื่องจากตำแหน่งของบุคคลจากภาพสีเกิดอาจจะมีหลายหลายก้อนวัตถุ (Blob) จึงต้องทำการเปรียบเทียบกับทุก ๆก้อนวัตถุในภาพสี โดยในขั้นตอนแรกจะหาค่าพื้นที่ส่วนที่ Overlap ระหว่างก้อนวัตถุที่เป็นตำแหน่งของบุคคลจากความลึกและสีโดยค่าเป็นเปอร์เซ็นต์เมื่อเทียบกับพื้นที่ของก้อนวัตถุที่เป็นตำแหน่งของบุคคลจากความลึก ($\cap_D$) จากนั้นจึงหาค่าความคาดเคลื่อนโดยรวม (ϵ) ของค่าตำแหน่งในแนวราบใหม่ ($\mathcal{X}_{new}$) กับ ค่าตำแหน่งในแนวราบของก้อนวัตถุบุคคลจากภาพสี ($\mathcal{X}_{rgb}$) และค่าขนาดความกว้างของตัวบุคคลใหม่ (w_{new}) กับ ค่าขนาดความกว้างของก้อนวัตถุบุคคลจากภาพสี (w_{rgb}) โดยให้ความสำคัญของค่าความคาดเคลื่อนของค่าตำแหน่งในแนวราบ และขนาดความกว้าง ตามลำดับดังนี้ $\gamma = 0.7$ และ $\delta = 0.3$ ซึ่งสามารถคำนวณค่าความคลาดเคลื่อนได้ตามสมการที่ (3.38)

$$\epsilon = \gamma|\mathcal{X}_{new} - \mathcal{X}_{rgb}| + \delta|w_{new} - w_{rgb}| \tag{3.38}$$

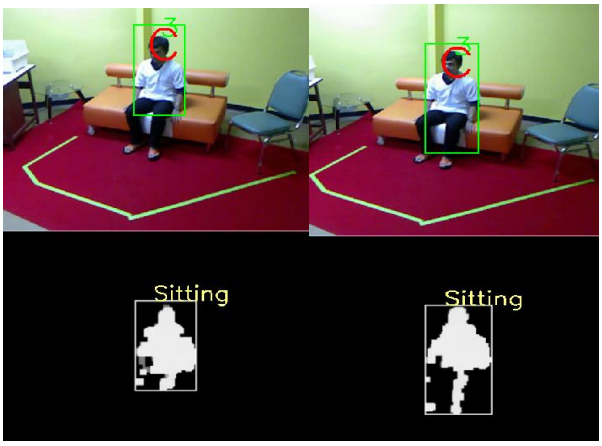
จากนั้นจะพิจารณาเฉพาะก้อนวัตถุบุคคลจากภาพสี ที่มีค่าพื้นที่ส่วนที่ Overlap ($\cap_D$) มากกว่า 75 เปอร์เซ็นต์ และค่าความคลาดเคลื่อน (ϵ) น้อยกว่า 50 หากเกินกว่านั้นถือว่าไม่มีก้อนวัตถุบุคคลจากภาพสีที่น่าเชื่อถือที่จะนำมาเปรียบเทียบเพื่อปรับค่าได้ ให้ใช้ค่าที่ได้จากการ Shift ค่าโดยตรงแทน ซึ่งถ้าหากมีก้อนวัตถุบุคคลจากภาพสี โดยเลือกข้อมูลจากก้อนวัตถุบุคคลจากภาพสีที่มีค่าความคลาดเคลื่อน (ϵ) ที่น้อยที่สุดมาอัพเดทตำแหน่งตามโดยใช้สัดส่วนตัวคูณจากค่าขนาดความกว้างของตัวบุคคลใหม่ (w_{new}) และค่าขนาดความกว้างของก้อนวัตถุบุคคลจากภาพสี (w_{rgb}) เพื่อนำไปใช้เป็นสัดส่วนตัวคูณเพื่อปรับค่าเช่นเดียวกับ Shift ค่าโดยตรง

(ข) การดึงคุณสมบัติเพื่อใช้ในการเปรียบเทียบเพื่อติดตามและจดจำตัวบุคคล (Comparison Properties for Tracking)

ในการติดตามและจดจำตัวบุคคลจำเป็นต้องมีคุณสมบัติที่ใช้ในการติดตามบุคคลที่กำลังอยู่ในเฟรมภาพ ซึ่งจะใช้คุณสมบัติของตำแหน่งในภาพในการติดตามเป็นหลัก และการจดจำตัวบุคคลจะใช้คุณสมบัติของตัวบ่งชี้สี (Color Descriptor) เป็นตัวอธิบายคุณลักษณะของแต่ละบุคคลเพื่อใช้ในการจดจำ

(1) ตำแหน่งของศีรษะ (Head of Position)

เนื่องจากใช้คุณสมบัติของกรอบตำแหน่งของบุคคล (Bounding Box) ในภาพความลึกมีความน่าเชื่อถือเหมาะแก่การในมาใช้ติดตามบุคคลที่กำลังอยู่ในเฟรมภาพ ซึ่งอธิบายโดยมีค่าตำแหน่งมุมซ้ายของบุคคลโดยบอกค่าในแนวราบและแนวดิ่ง เป็นคู่อันดับ X และY และบอกขอบเขตของวัตถุโดยใช้ค่าความกว้าง และความสูงตาม Bounding Box ของวัตถุ ดังภาพประกอบที่ 3-12 ซึ่งเมื่อพิจารณาแล้วคุณสมบัติและตำแหน่งของตัวบุคคลที่จะนำมาติดตามควรจะเป็นจุดศูนย์กลาง แต่เนื่องด้วยค่าจุดศูนย์กลางต้องอาศัยคุณสมบัติการคำนวณจากค่าความสูงซึ่งมีความแปรปรวนสูงเนื่องจากตำแหน่งของเท้าและส่วนล่างอื่น ๆจะใกล้เคียงกับพื้นจึงทำให้ในบางเฟรมไม่สามารถจะตรวจจับได้ครบและมีค่าแกว่งไปมา ตามตัวอย่างท่านั่ง ภาพประกอบที่ 3-15 จึงจำทำให้ต้องใช้บริเวณตำแหน่งของศีรษะมาใช้ในการติดตามแทน



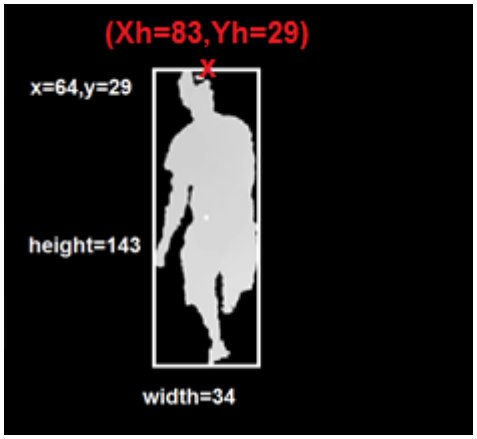
ภาพประกอบที่ 3-15 ความแปรปรวนสูงในการตรวจจับตำแหน่งของขา

สำหรับการคำนวณค่าตำแหน่งของศีรษะ จะใช้ตำแหน่งตรงกลางบนสุดของศีรษะมาใช้ในการติดตามตำแหน่งของศีรษะ ซึ่งจะคำนวณจากกรอบตำแหน่งของบุคคล (Bounding Box) ในภาพความลึก เป็นคู่อันดับ X ($\mathcal{X}_d$) และY (Y_d) และบอกขอบเขตของวัตถุโดยใช้ค่าความกว้าง (W_d) โดยคำนวณตามสมการที่ (3.39) และ (3.40)

$$\mathcal{X}_h = \mathcal{X}_d + (W_d/2) \tag{3.39}$$

$$Y_h = Y_d \tag{3.40}$$

ซึ่งจะได้ตำแหน่งในแนวราบของศีรษะ (X_h) และตำแหน่งในแนวดิ่งของศีรษะ (Y_h) ตามภาพประกอบที่ 3-16

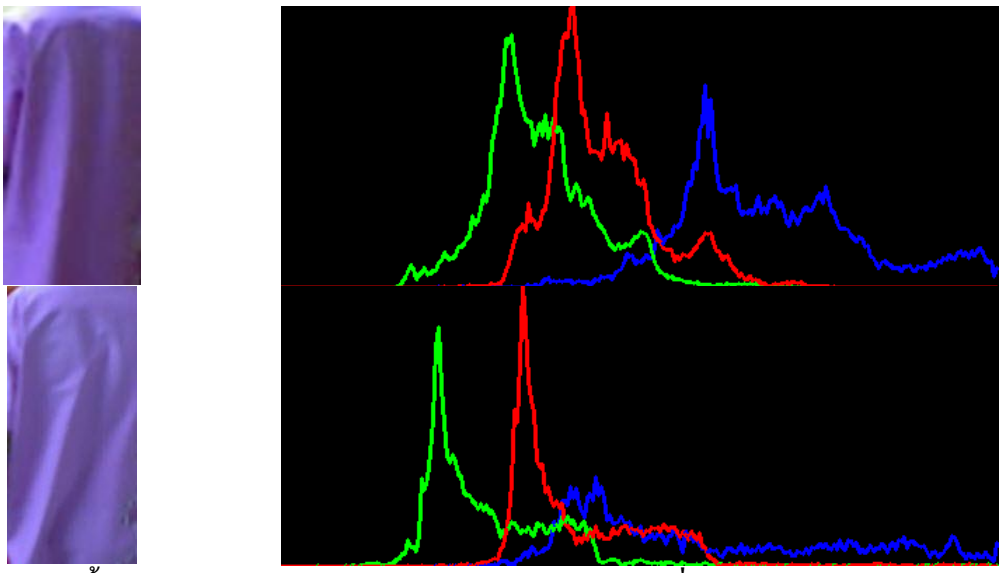


ภาพประกอบที่ 3-16 ตำแหน่งในแนวราบของศีรษะ (X_h) และตำแหน่งในแนวดิ่งของศีรษะ (Y_h)

(2) คุณสมบัติของตัวบ่งชี้สี (Color Descriptor)

ในงานวิจัยชิ้นนี้จะใช้ภาพของบุคคลที่มีสีในโหมด RGB คือ Red Green Blue ซึ่งเกิดจากการผสมระหว่าง 3 แสงสีในสัดส่วนความเข้มข้นที่แตกต่างกัน เมื่อนำมาผสมกันทำให้เกิดสีต่าง ๆ บนจอคอมพิวเตอร์ได้มากถึง 16.7 ล้านสี ซึ่งใกล้เคียงกับสีที่ตาผู้วิจัยมองเห็นได้โดยปกติ และจุดที่สีทั้งสามสีรวมกันจะกลายเป็นสีขาวที่เท่ากัน นิยมเรียกการผสมสีแบบนี้ว่าแบบ “Additive” หรือการผสมสีแบบบวก และเมื่อเปลี่ยนอัตราส่วนต่าง ๆจะสามารถทำให้เกิดโทนสีขึ้นมากมาย

แต่ในโหมดสี RGB แม้ว่าจะบุคคลจะมีรูปลักษณ์เป็นสารสีเดียว แต่ก็ยังมีความแปรปรวนไปตามสภาพแสงที่กระทบกับบุคคลจึงทำให้ไม่สามารถนำโหมดสีใน RGB มาจดจำบุคคลได้ ดังภาพประกอบที่ 3-17 ซึ่งตัวอย่างที่แสดงภาพสีเสื้อของบุคคลที่ต่างบริเวณกัน 3-4 เมตร ที่มีแสงสว่างไม่เท่ากัน จะทำให้มีค่าสีซึ่งแสดงตามได้กราฟความถี่สะสม (Histogram)



(ก) ภาพสีเสื้อตัวเดียวกันในสภาพแสงต่างกัน (ข) กราฟสะสมความถี่ (RGB Histogram)

ภาพประกอบที่ 3-17 ความแตกต่างของภาพสีเสื้อตัวเดียวกันในสภาพแสงต่างกัน

ดังนั้น ในงานวิจัยชิ้นนี้จึงต้องใช้การนำเสนอข้อมูลแบบใหม่ที่สามารถบีบข้อมูลช่วงสี และดึงลักษณะสีที่เด่นเมื่อเทียบกับกับสีทั้งหมดเพื่อใช้ในการรู้จำ และอยู่ในข้อกำหนดที่สามารถทำได้อย่างรวดเร็วอย่างทันเวลา (Real-time Condition) ในงานวิจัยนี้จึงนำเสนอวิธีการแปลงภาพให้อยู่ในโหมด RG Chromaticity ซึ่งเป็น Space ในการนำเสนอข้อมูล สองมิติที่นอมัลไลท์สีใน RGB ซึ่งก็คือ “Chromaticity Space” ซึ่งโดยทั่วไปของ RGB Color Space ในแต่ละพิกเซลจะมีเอกลักษณ์ตามระดับความเข้มข้นของสีแดง เขียว และฟ้า เป็นสีหลัก ดังนั้น สีแดงสดที่สว่างสามารถจะแสดงได้เป็น (R,G,B) มีค่า (255,0,0) ในขณะที่ สีแดงเข้มซึ่งมืดจะมีค่าอยู่ในช่วง (40,0,0) ในการนอมัลไลท์ RGB Color Space หรือ RG Color Space โดยที่จะถูกจัดแสดงใหม่อยู่ในรูปของความเป็นสัดส่วนของทั้ง สีแดง สีเขียว และสีฟ้า เมื่อเปรียบเทียบกับสัดส่วนความเข้มของแต่ละสี ซึ่งเมื่อรวมค่าของ Chromaticity ทั้งสามสีแล้วจะได้เท่ากับ 1 กล่าวคือ จะนำสีนั้นมาเทียบกับค่าของสีทั้งสามที่รวมกัน โดยปกติแล้วงานวิจัยอื่น ๆจะไม่นำสีฟ้ามาใช้งาน แต่สามารถนำมาคำนวณได้ ซึ่งในงานวิจัยนี้จะสนใจค่าของ Chromaticity ทั้งสามสี ซึ่งทำให้เกิดเป็น RGB Chromaticity Space

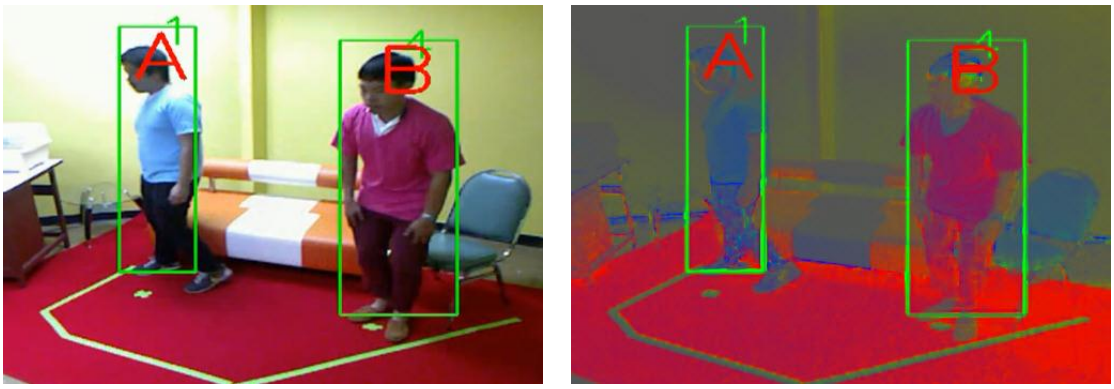
สำหรับการเปลี่ยน RGB Color Space ไปยัง RGB Chromaticity Space จะทำได้โดยดึงค่าความเข้ม (Intensity) ในโหมดสี RGB และคำนวณอัตราส่วนออกมาเป็น Chromaticity ของ R G B และคูณด้วยค่า Bit ของสีที่ใช้แสดงกลับไป โดยในงานวิจัยนี้จะใช้ระดับสีที่ 8 bits มีค่าสูงสุดที่ Mr=255 ดังสมการที่ (3.41), (3.42) และ (3.43) ตามลำดับ

$$r = \frac{R}{R+G+B}\ x\ Mr \qquad (3.41)$$

$$g = \frac{G}{R+G+B} x\ Mr \tag{3.42}$$

$$b = \frac{B}{R+G+B} x\ Mr \tag{3.43}$$

ซึ่งเมื่อคำนวณออกมาแล้วค่า $r + g + b$ จะเท่ากับ Mr โดยแสดงตัวอย่างของการแปลง RGB Color Space ไปยัง RGB Chromaticity Space ดังภาพประกอบที่ 3-18

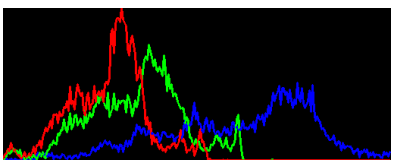
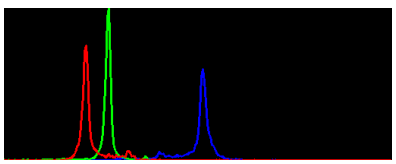
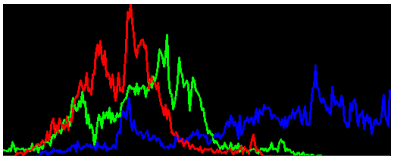
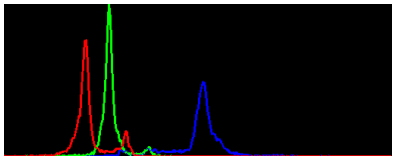


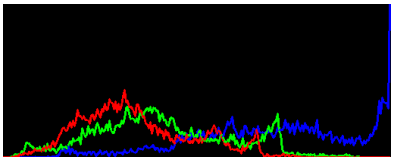
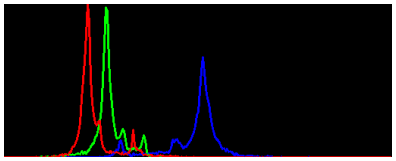
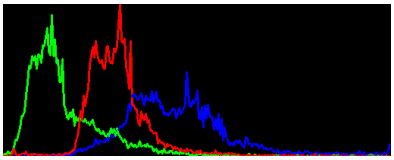
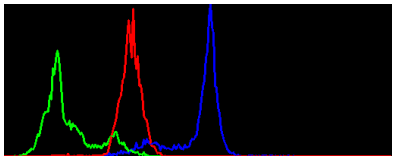
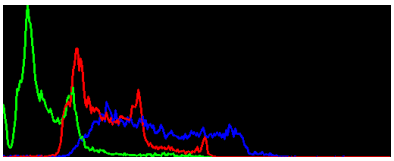
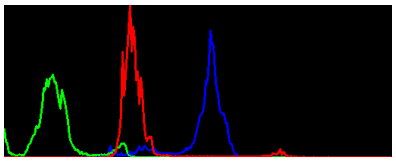
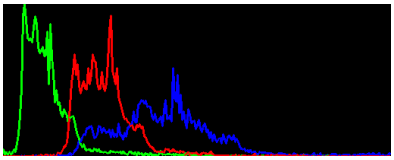
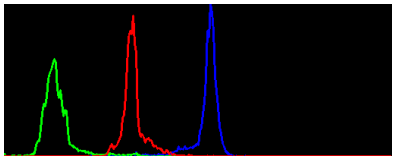
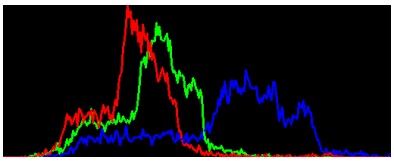
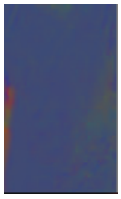
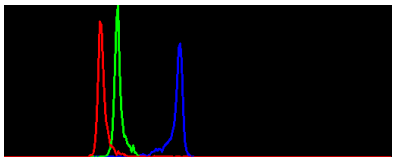
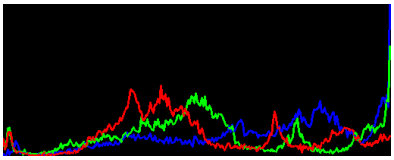
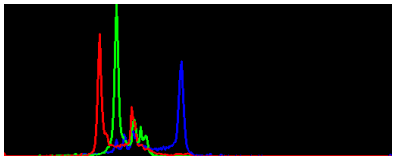
(ก) RGB Color Space (ข) RGB Chromaticity Color Space

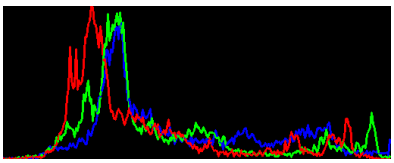
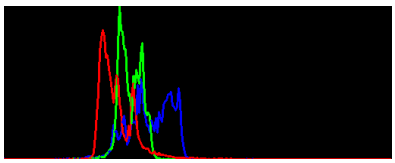
ภาพประกอบที่ 3-18 ตัวอย่างภาพ RGB Color Space และ RGB chromaticity space

ซึ่งจากผลการทดลองการแปลงภาพเป็น RGB Chromaticity Space จากเสื้อจะทำให้ค่าสีถูกบีบให้อยู่ในช่วงแคบ ที่มีความแตกต่างกันในแต่ละสี แต่จะคงลักษณะใกล้เคียงมากขึ้นเมื่อเป็นสีเสื้อเดียวกันที่ต่างกันที่สภาพแสง ซึ่งแสดงตัวอย่างการทดลองได้ดังตารางที่ 3-5

ตารางที่ 3-5 ตัวอย่างผลการทดลองการแปลงภาพเป็น RGB Chromaticity Space จากเสื้อ

Object	Histogram of Object	Chromatic Image	Histogram of Chromatic Image
A1 			
A2 			

Object	Histogram of Object	Chromatic Image	Histogram of Chromatic Image
A3			
B1			
B2			
B3			
C1			
C2			

Object	Histogram of Object	Chromatic Image	Histogram of Chromatic Image
C3 			

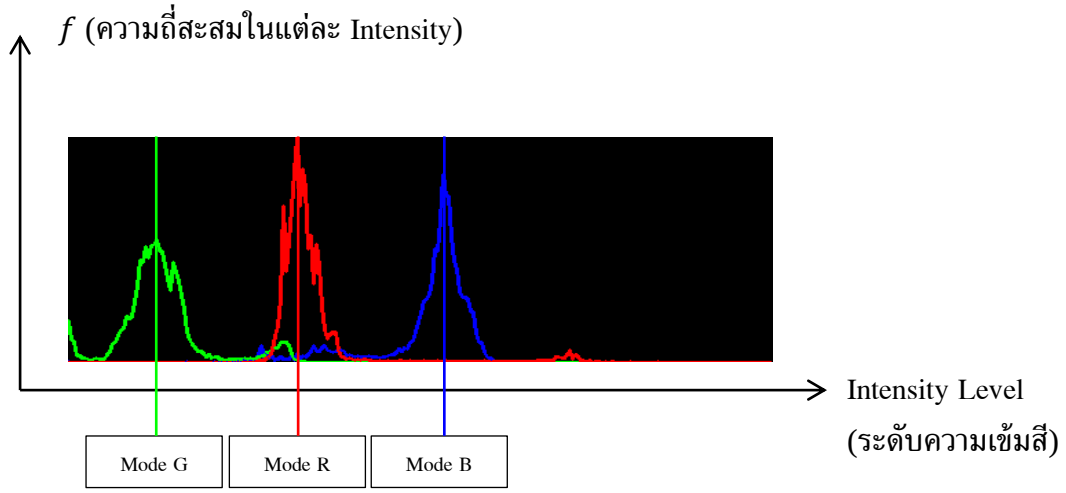
ขั้นตอนถัดไปจะเป็นการหาบริเวณที่จะดึงสีขึ้นมา เนื่องจากภาพที่ผู้วิจัยได้เป็นกรอบสี่เหลี่ยม ซึ่งจะติดพื้นหลังมาด้วย จึงจำเป็นต้องดึงเฉพาะส่วนที่เป็นตัวบุคคลขึ้นมาเท่านั้นเพื่อจะได้สีที่เป็นสีจริงของบุคคลขึ้นมาเท่านั้น โดยการใช้ตำแหน่งที่ซ้อนทับกันของภาพความลึกที่ถูกขีดแบ่งค่าตามเกณฑ์ เป็นไบนารี่แล้ว โดยมีค่าเป็น 0 และค่าสูงสูดของสี 255 มาทำกระทำกันโดยใช้ AND Operator ทั้งสาม Channel R G และ B ซึ่งจะได้ผลลัพธ์ดังตัวอย่าง ตามภาพประกอบที่ 3-19

(ก) ภาพสีของบุคคล (ข) ภาพความลึกของบุคคล (ค) ภาพที่นำพื้นหลังออก

ภาพประกอบที่ 3-19 ตัวอย่างการทำการตัดพื้นหลังของตัวบุคคลในภาพสีออก

ซึ่งจากผลลัพธ์พื้นหลังจะกลายเป็นสีดำ ซึ่งสามารถหักล้างเมื่อนำภาพที่ทำการตัดพื้นหลังของตัวบุคคลในภาพสีออกแล้วนำไปทำเป็น Histogram แล้วจึงนับพิกเซลที่เป็นสีดำไปหักล้างออกหลังจากทำ Histogram ได้ในภายหลัง

เมื่อหาค่า Histogram ได้แล้ว ขั้นตอนต่อไปก็คือการดึงคุณสมบัติของค่าสีในบุคคลต่าง ๆ โดยในงานวิจัยนี้สนใน ค่าฐานนิยม (Mode) ซึ่งเป็นค่าที่มีค่าซ้ำกันมากที่สุดในภาพ ซึ่งเทียบได้กับค่า Intensity ที่มีความถี่มากที่สุดใน Histogram ของแต่ละ Channel สี ซึ่งจะได้ Mode ของ r (m_r) , Mode ของ g (m_g) และ Mode ของ b (m_b) ดังแสดงตัวอย่างในภาพประกอบที่ 3-20



ภาพประกอบที่ 3-20 ตัวอย่างการหาค่าในค่าฐานนิยม (Mode) ในแต่ละ Channel สี

จากนั้นทำการหาค่าของเนื้อสีของแต่ละสีโดยเป็นผลรวมจาก Histogram ในแต่ละ Channel ของสี ตามสมการที่ (3.44) , (3.45) และ (3.46) แล้วรวมค่าของเนื้อสีทั้งหมด ตามสมการที่ (3.47)

$$S_r = \sum_{i=0}^{\max(i)} H_r(i) \tag{3.44}$$

$$S_g = \sum_{i=0}^{\max(i)} H_g(i) \tag{3.45}$$

$$S_b = \sum_{i=0}^{\max(i)} H_b(i) \tag{3.46}$$

$$S_{all} = S_r + S_g + S_b \tag{3.47}$$

หลังจากนั้นทำการ Normalization เพื่อหาสัดส่วนของแต่ละสีทั้งหมดออกมาเป็นเปอร์เซ็นต์ ก็จะได้ค่าสัดส่วนของเนื้อสีแดง (ρ_r), สัดส่วนของเนื้อสีเขียว (ρ_g), และสัดส่วนของเนื้อสีน้ำเงิน (ρ_b) ตามสมการที่ (3.48), (3.49) และ (3.50)

$$\rho_r = \left(\frac{S_r}{S_{all}}\right) x100 \tag{3.48}$$

$$\rho_g = \left(\frac{S_g}{S_{all}}\right) x100 \tag{3.49}$$

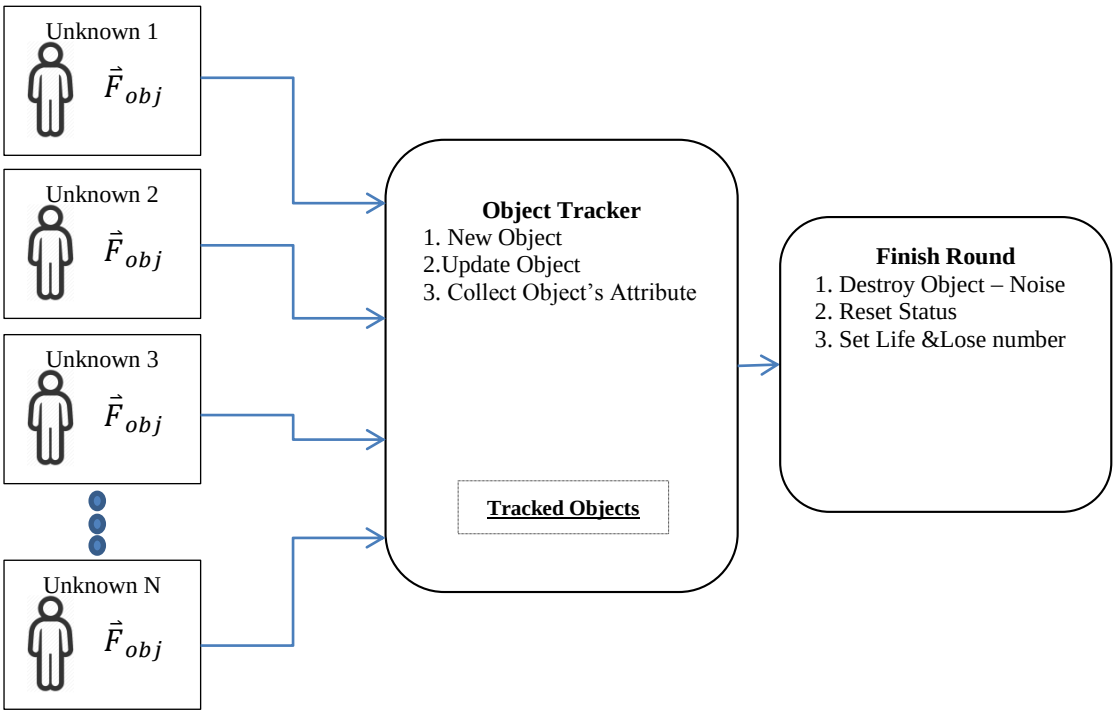
$$\rho_b = \left(\frac{S_b}{S_{all}}\right) x100 \qquad (3.50)$$

เมื่อหาค่าทั้งหมดแล้วก็จะได้ค่า Feature Set ของตัวบ่งชี้ค่าสี (Color Descriptor) ที่ประกอบไปด้วยค่าฐานนิยม (Mode) ในแต่ละ Channel ของสี และค่าสัดส่วนเนื้อสีของแต่ละสี รวมไปถึงข้อมูลที่เป็นตำแหน่งของบุคคลซึ่งจะแทนด้วยตำแหน่งในแนวราบของศีรษะ (X_h) และตำแหน่งในแนวดิ่งของศีรษะ (Y_h) ก็จะได้ Feature Set ดังสมการที่ (3.51) เพื่อใช้ในการติดตามและจดจำบุคคลต่อไป

$$\vec{F}_{obj} \in \{X_h, Y_h, m_r, m_g, m_b, \rho_r, \rho_g, \rho_b\} \qquad (3.51)$$

(ค) การติดตามและจดจำตัวบุคคลในมุมมองเดี่ยว (Object Tracking and Re-Identifying)

สำหรับส่วนในการติดตามและจดจำบุคคลจะรับข้อมูลของแต่ละคนซึ่งมีข้อมูลทั้งตำแหน่ง (Head Position) และตัวบ่งชีค่าสี (Color Descriptor) เพื่อที่นำไปเปรียบเทียบกับบุคคลที่กำลังติดตามอยู่ทั้งที่ออกนอกเฟรมไปแล้ว ทั้งที่อยู่ในเฟรม หากเปรียบเทียบกับบุคคลที่ถูกติดตามในระบบแล้วไม่มีก็จะทำการสร้างขึ้นมาใหม่ (New Object) หากเปรียบตรงกัน ก็จะทำการปรับปรุงข้อมูลให้เป็นปัจจุบัน (Update Object) และสุดท้ายจะเป็นการเก็บข้อมูลของ Object นั้น ที่ใช้ในรู้จำท่าทางและอื่น ๆ เข้าไปเก็บในข้อมูลของบุคคลที่ระบบกำลังติดตามอยู่ (Collect Object's Attribute) เมื่อทุกบุคคลที่ถูกตรวจจับได้ถูกประมวลผลโดยตัวติดตาม ระบบก็จะเรียกฟังก์ชัน Finish Round เพื่อทำการทำลาย Object ที่เป็นสิ่งรบกวนออก, เปลี่ยนสถานะบาง Object ที่ไม่ถูกพบในเฟรมปัจจุบัน เป็น Inactive และ Lost รวมไปถึงตั้งค่าเวลาการอยู่ในเฟรม Life Number และเวลาที่หายไปจากเฟรม Lose Number ซึ่งเป็นตัวที่จำเป็นสำหรับการตั้งค่าสถานะต่าง ๆของวัตถุที่ถูกติดตาม (ในงานวิจัยนี้จะเรียกแทนบุคคลว่าวัตถุ) ดังที่แสดงในภาพประกอบที่ 3-21



ภาพประกอบที่ 3-21 ภาพรวมการทำงานของระบบติดตามในมุมมองเดี่ยว

(1) การเปรียบเทียบค่า Error ระหว่าง Unknown Object และTracked Objected (Error – Comparison)

ก่อนการเปรียบเทียบค่า Error วัตถุที่อยู่ในระบบทุกชิ้น จะเก็บข้อมูลของ Feature Set ($\vec{F}_{obj}$) ก่อนหน้าและที่เข้ามาใหม่เพื่อเป็นประวัติไว้ โดยมีค่าสูงสุดเป็นจำนวนเท่ากับ $NH_{\vec{F}}$ เพื่อนำมาเฉลี่ย และหาค่ากลางเป็นจำนวน $CH_{\vec{F}}$ ซึ่งอาจจะมีค่าเท่ากับ $NH_{\vec{F}}$ ก็ได้ ซึ่งจะถูกนำมาเฉลี่ยตามแต่ละชนิดของ Feature ซึ่งจะเป็นตัวแทนของวัตถุที่เคยเจอในระบบซึ่งแทนด้วย $\vec{F}_{avg}$

โดยในระบบจะมีค่า Error อยู่ 2 ประเภท ตามลักษณะของ Feature คือ Distance Error และ Color Distance โดยจะเปรียบเทียบตามคู่ของแต่ละตัวระหว่าง $\vec{F}_{avg}$ และ $\vec{F}_{obj}$ (Unknown Object) โดยสามารถหาค่าของ Distance Error (ε_d) ตามสมการที่ (3.52) และค่าของ Color Error (ε_c) ตามสมการที่ (3.53)

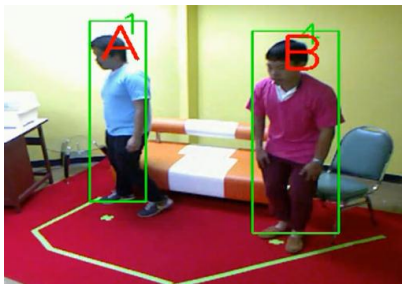
$$\varepsilon_d = \sqrt{\left(\vec{F}_{obj}[\mathcal{X}_h] - \vec{F}_{avg}[\mathcal{X}_h]\right)^2 + \left(\vec{F}_{obj}[Y_h] - \vec{F}_{avg}[Y_h]\right)^2} \qquad (3.52)$$

$$\varepsilon_c = \sqrt{\sum_{c=r}^{b}\left(\vec{F}_{obj}\,[\mathrm{m}(c)] - \vec{F}_{avg}\,[\mathrm{m}(c)]\right)^2 + \sum_{c=r}^{b}\left(\vec{F}_{obj}\,[\rho(c)] - \vec{F}_{avg}\,[\rho(c)]\right)^2} \qquad (3.53)$$

(2) สถานะของวัตถุในระบบ (Object Statuses of System)

ในงานวิจัยนี้จะแบ่งสถานะของวัตถุที่ถูกติดตามในระบบเป็น 5 สถานะที่แตกต่างกัน โดยระบบจะมีตัวเก็บค่า Lose Number ซึ่งบ่งบอกว่าวัตถุที่เคยถูกตรวจพบนั้นหายไปจากระบบนานเท่าไหร่ และ Life Number คือ ค่าที่บอกว่าวัตถุเข้ามาในเฟรมนานเท่าไหร่ตั้งแต่เริ่มและไม่ลบล้างเมื่อวัตถุหายไป ซึ่งค่า Lose Number จะมีบทบาทมากในการเปลี่ยนสถานะของวัตถุที่อยู่ในระบบ โดยค่าเหล่านี้จะถูกตั้งค่าใหม่เมื่อประมวลผลทุกวัตถุที่เข้ามาในเฟรมเสร็จสิ้น และเข้าสู่กระบวนการ END Round ซึ่งวัตถุที่เคยถูกติดตามในระบบไม่ถูก Match ก็จะไม่ถูก Stamp Round ID ซึ่งทำให้รู้ว่าวัตถุใดหายไป หรือวัตถุไหนยังถูกติดตามอยู่ หรือวัตถุไหนหายไปแล้วกลับมา ซึ่งระบบก็จะทำการตั้งค่า Lose Number และ Life Number รวมไปถึง Reset สถานะของวัตถุต่าง ๆในระบบ ซึ่งสถานะของวัตถุที่ถูกติดตามในระบบเป็น 5 สถานะ ดังนี้

- **Active Object** คือ วัตถุที่ปรากฏ ณ ปัจจุบัน หรือหายไปเล็กน้อยอาจจะ Lose Number เป็น 1-5 เฟรม ซึ่งอาจจะเคยเป็น Inactive หรือ Initial มาก่อน จะแสดงเป็นสีเขียว ดังภาพประกอบที่ 3-22



ภาพประกอบที่ 3-22 ตัวอย่าง Active Object ที่แสดงโดยกรอบสีเขียว

- **Inactive Object** คือ วัตถุที่เคยเข้ามาในระบบ แต่ปัจจุบันไม่อยู่ในเฟรม
- **Initial Object** คือ วัตถุที่ถูกค้นพบใหม่ที่ยังไม่ถูกตรวจพบว่าเป็น Active หรือมีความคล้ายกับ Inactive อันเนื่องมาจากวัตถุยังมี Feature ที่ไม่คงที่ ต้องใช้เวลาเพื่ออัพเดทColor Descriptor ให้ใกล้เคียงกับ Inactive/Active Object ซึ่งเคยตรวจจับได้หรืออาจจะเป็นวัตถุที่เป็นบุคคลใหม่ซึ่งไม่เคยเข้ามาในระบบ โดยที่ Initial Object เมื่อถูกอัพเดทตามเวลาที่กำหนดก็จะถูกนำไปเปรียบเทียบอีกครั้งกับ Inactive/Active Object หากยังไม่ใช่ก็จะถือว่าเป็นบุคคลใหม่ซึ่งไม่เคยเข้ามาในระบบ ซึ่ง Initial Object จะแสดงเป็นสีขาว ดังภาพประกอบที่ 3-23

ภาพประกอบที่ 3-23 ตัวอย่าง Initial Object ที่แสดงโดยกรอบสีขาว

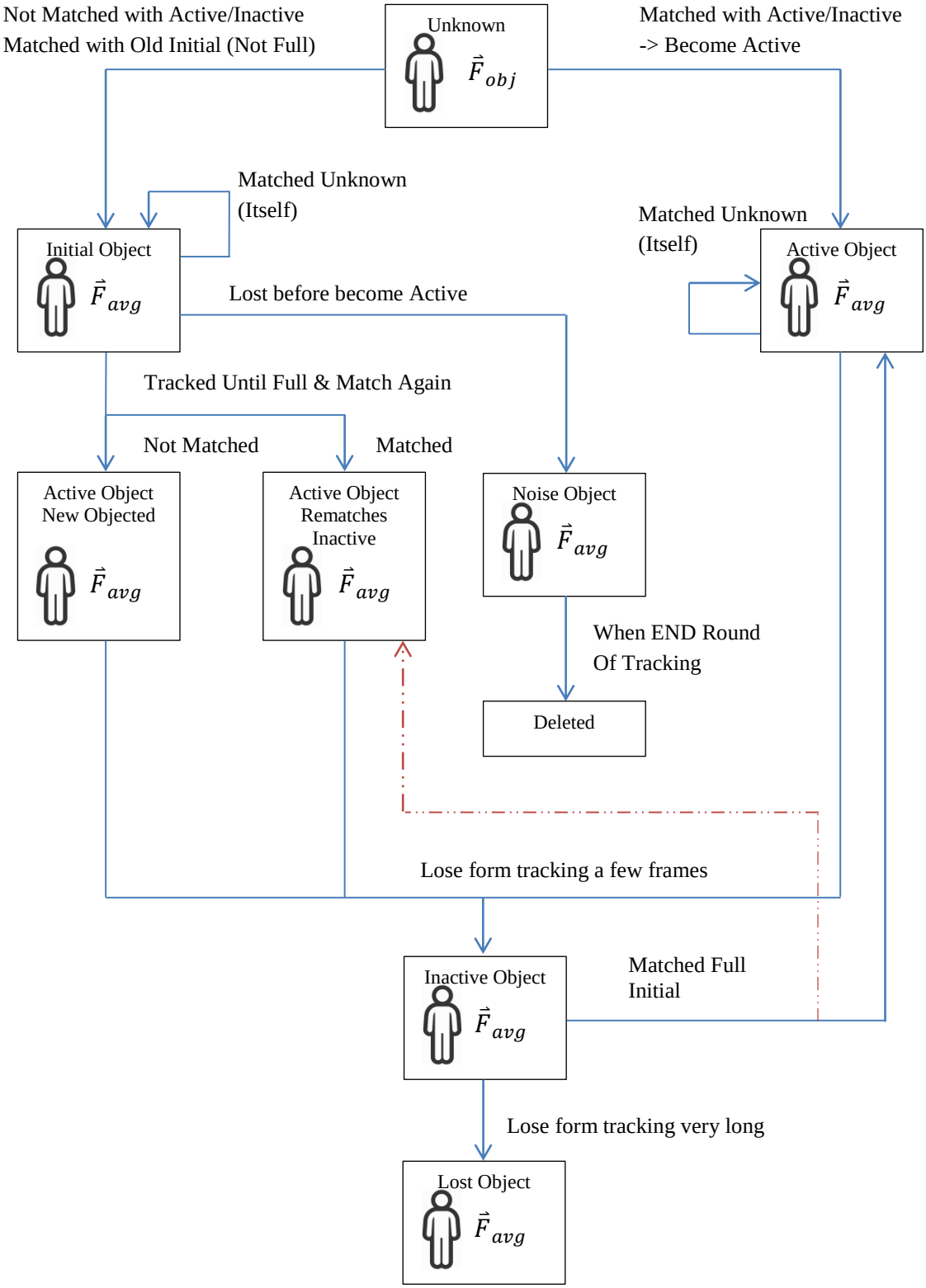
- **Noise Object** คือ วัตถุที่หายไปจากเฟรมนาน ซึ่งมีค่า Lose Number สูง และเคยมีสถานะเป็น Initial มาก่อน ซึ่งอาจเกิดจากการติดตามไม่ทัน หรืออาจเกิดจากสิ่งรบกวนของภาพความลึก
- **Lost Object** คือ วัตถุที่เคยมีสถานะเป็น Inactive Object และหายไปนานเกินค่าที่ระบบตั้งไว้

ตารางที่ 3-6 สรุปสถานะของวัตถุที่กำลังติดตาม

สถานะ	**คำอธิบาย**	**สถานะก่อนหน้า**	**สถานะถัดไป**	**ค่าที่ใช้ในการเปรียบติดตามสถานะเดิม**
Active	เป็นวัตถุที่กำลังถูกติดตาม ณ ปัจจุบัน	*Initial*, *Inactive*	เมื่อหายไปช่วงหนึ่ง จะเป็น *Inactive*	*Distance Error* ช่วงกลาง & *Color Error* ช่วงกว้าง
Inactive	เป็นวัตถุที่เคยถูกตรวจจับเป็น *Active* ในระบบแต่ไม่ปรากฏในปัจจุบัน	*Active*	เมื่อหายไปนาน เช่น 3 ชม. หรือตามที่ระบบตั้งไว้ จะกลายเป็น *Lost*	ไม่มี
Initial	เป็นสถานะเริ่มต้นที่ไม่ถูกพบว่าเป็นวัตถุที่เคยเข้ามาในระบบ จะให้เป็น *Initial* ก่อนแล้วจะมีการพิจารณาอีกครั้ง	ไม่มี	*Active(New Object)*, *Active (OLD from Inactive)*	*Distance Error* ช่วงแคบ

สถานะ	คำอธิบาย	สถานะก่อนหน้า	สถานะถัดไป	ค่าที่ใช้ในการเปรียบติดตามสถานะเดิม
Noise	เป็นวัตถุที่เป็นขยะของระบบ ซึ่งเป็น *Initial* ที่ปรากฏขึ้นไม่นาน หายไปนานเกินกว่าที่ระบบกำหนด	*Initial*	รอการทำลาย *Destroy* เมื่อ *End Round*	ไม่มี
Lost	วัตถุที่เคยมีสถานะเป็น Inactive Object และหายไปนานเกินค่าที่ระบบตั้งไว้	*Inactive*	ไม่มี แต่ไม่ถูกทำลาย ยังมีการเก็บข้อมูลไว้เพื่อตรวจสอบ	ไม่มี

ซึ่งโดยสรุปแล้ววัตถุที่เข้ามาใหม่ที่ระบบไม่รู้จัก (Match ไม่ตรงกับ Active และ Inactive ในระบบ) จะเริ่มเป็น Initial ก่อน เมื่อ Initial ถูก Track และอัพเดทจนครบระยะเวลาที่กำหนด จะถูกพิจารณาอีกครั้งโดยการเปรียบเทียบว่ามีคุณสมบัติตรงกับ Active และ Inactive หรือไม่ หากตรงก็ทำการอัพเดทและฟื้นคืนสถานะ หากไม่ตรงก็จะทำการสร้างวัตถุขึ้นมาใหม่เป็น Active ต่อมาหาก Active หายไปจากเฟรมระยะหนึ่ง เช่น 3-5 เฟรม ซึ่งแล้วแต่จะกำหนด จะถือว่าวัตถุตัวนั้นหายไปจากระบบก็จะกลายสถานะเป็น Inactive และหาก Inactive หายไปนานมากก็จะถือว่าเป็น Lost และไม่นำกลับมา Match อีก แต่ข้อมูลยังคงเก็บไว้ในระบบอยู่ ซึ่งสามารถอธิบายเป็นวงจรชีวิตของ Object ตามภาพประกอบที่ 3-24



ภาพประกอบที่ 3-24 วงจรชีวิตและสถานการณ์เปลี่ยนแปลงของวัตถุที่อยู่ในระบบ

(3) การประมวลผลวัตถุหนึ่ง ๆที่เข้ามาในเฟรมภาพ (Processing in an Unknown Object)

การประมวลผลกับ Unknown Object ที่เข้ามาใหม่ที่อยู่ในเฟรมกับวัตถุที่มีอยู่ในระบบทั้ง Initial, Active และ Inactive จะต้องหาค่า Error วัตถุที่มีอยู่ในระบบทุกชิ้นและการใช้เกณฑ์เพื่อเปรียบเทียบในการตัดสินใจว่าวัตถุนั้นเป็นวัตถุที่อยู่ในระบบหรือไม่

โดยในขั้นแรกระบบจะเปรียบเทียบวัตถุที่เข้ามาใหม่กับวัตถุชุดหนึ่ง ๆ (Initial หรือ Active หรือ Inactive) โดยหาค่า Error ของวัตถุที่เข้ามาใหม่กับวัตถุในชุดนั้นทุกตัว และเลือกค่าที่มี Error ที่น้อยที่สุด โดยจะมีการใช้ค่าที่เรียกว่า ขีดแบ่ง (Threshold) เป็นเกณฑ์ ซึ่งถ้าเปรียบเทียบกับค่าที่มี Error ที่น้อยที่สุดแล้วไม่เกินค่าขีดแบ่งที่ตั้งไว้ก็จะถือว่า Matched กับวัตถุที่มีค่า Error ที่น้อยที่สุด หากเกินค่าที่ตั้งไว้ก็จะถือว่า Miss Matched กับวัตถุชุดนั้น ๆ โดยการ Match กับวัตถุชุดหนึ่ง ๆนั้น (Initial หรือ Active หรือ Inactive) จะใช้ค่าขีดแบ่งที่ต่างกันทั้งชนิดของค่า Error และค่า Threshold ดังที่อธิบายตามตารางที่ 3-7

ตารางที่ 3-7 สรุปการ Match ที่ใช้ค่า Threshold ที่ต่างกันทั้งชนิดของค่า Error และค่า Threshold

สถานะที่ Match ด้วย	**ชนิดของค่า Threshold**	**ช่วงค่าโดยประมาณ**	**เลือกค่าที่น้อยที่สุดจาก**
Initial	Distance Threshold of Initial ($Ti\varepsilon_d$)	Distance Error (ε_d) ช่วงกลางประมาณ 50	Distance Error (ε_d)
Active	Distance Threshold of Active ($Ta\varepsilon_d$) Color Threshold of Active ($Ta\varepsilon_c$)	Distance Error (ε_d) ช่วงกลาง-กว้าง ประมาณ 50-80 และ Color Error (ε_c) ช่วงกว้าง 100-200	Color Error (ε_c)
Inactive	Color Threshold of Inactive ($Tia\varepsilon_c$)	Color Error (ε_c) ช่วงแคบ ประมาณ 17-30	Color Error (ε_c)

โดยสรุปแล้ว ในระบบจะมีค่า Threshold 4 ค่า ที่ใช้ต่างวาระกัน แล้วแต่ว่า Match กับชุดของ Object ใด ดังนี้

-Match กับ Initial

Distance Threshold of Initial ($Ti\varepsilon_d$)

-Match กับ Active

Distance Threshold of Active ($Ta\varepsilon_d$)

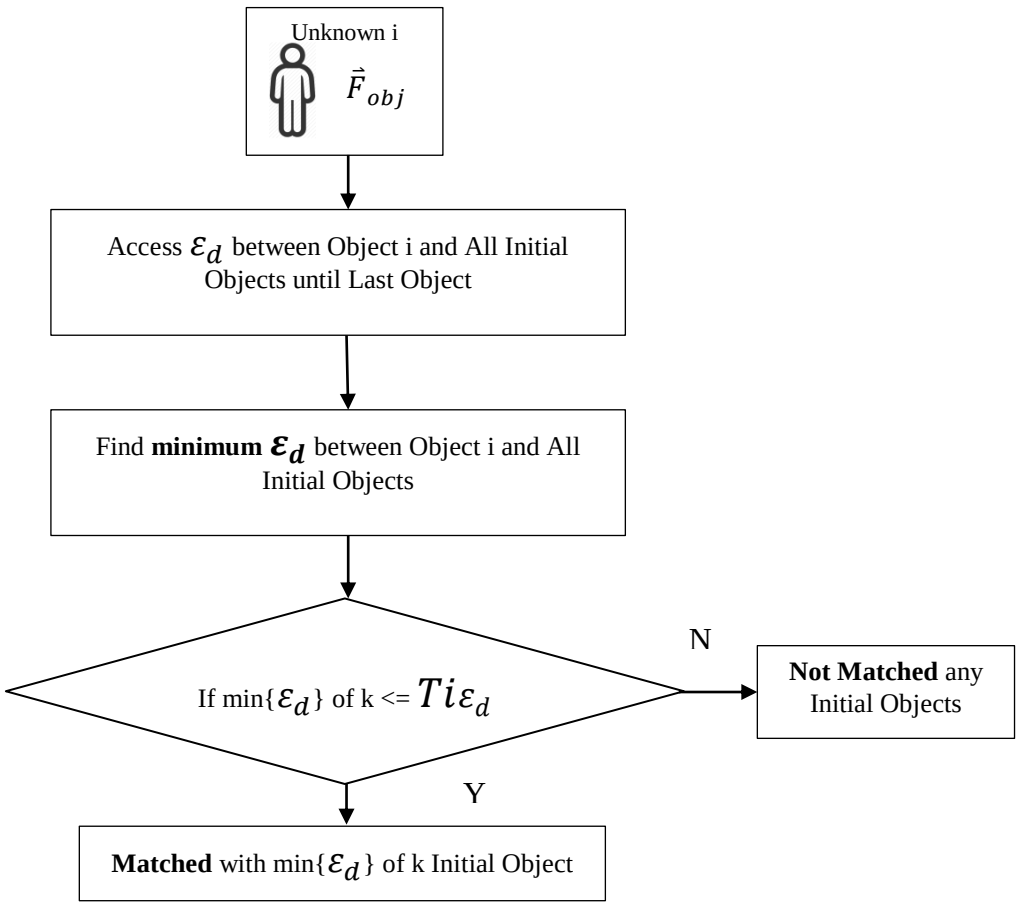
Color Threshold of Active ($Ta\varepsilon_c$)

-Match กับ Inactive

Color Threshold of Inactive ($Tia\varepsilon_c$)

(4) การ Match Unknown Object i กับ Initial Objects

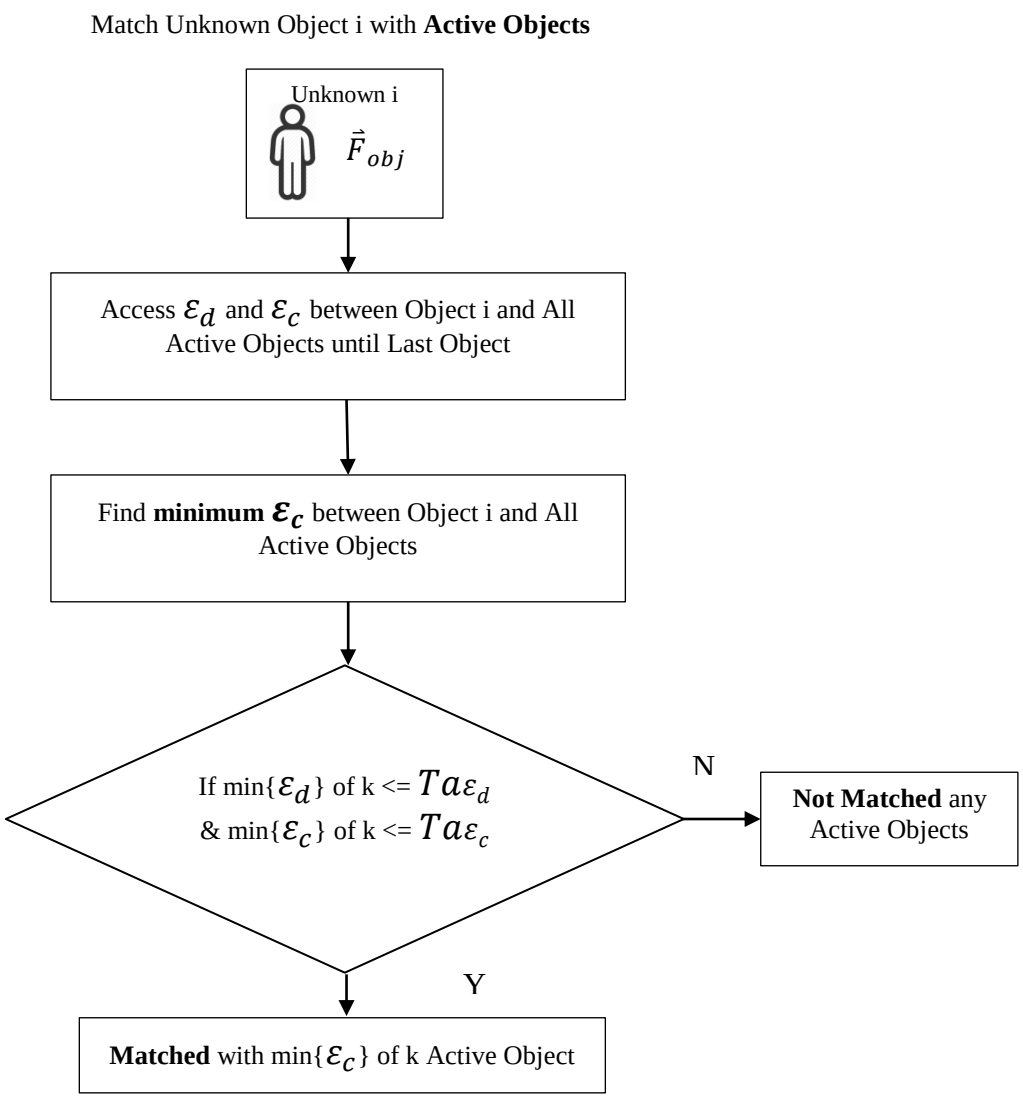
เนื่องจากวัตถุที่เข้ามาในเฟรมแรกเริ่มข้อมูลยังอาจจะไม่ครบทุกส่วน จึงมีค่าความแปรปรวนสูงมาก จำเป็นต้องติดตาม Initial จากค่า Distance Error (ε_d) โดยขั้นแรกระบบจะดึงค่า Distance Error ระหว่าง Object i กับทุก Initial Object ที่คำนวณไว้แล้วในระบบ และนำมาหาค่า Minimum ก็จะได้ค่าน้อยที่สุด min{ε_d} of k ที่เป็น Initial Object จากนั้นใช้เกณฑ์เพื่อเปรียบเทียบในการตัดสินใจว่าวัตถุนั้นควรเป็นวัตถุ k หรือไม่ ตามค่าขีดแบ่ง Distance Threshold of Initial ($Ti\varepsilon_d$) ซึ่งแสดงตามภาพประกอบที่ 3-25



ภาพประกอบที่ 3-25 วิธีการ Match Unknown Object i และ Initial Objects

(5) การ *Match Unknown Object i* กับ *Active Objects*

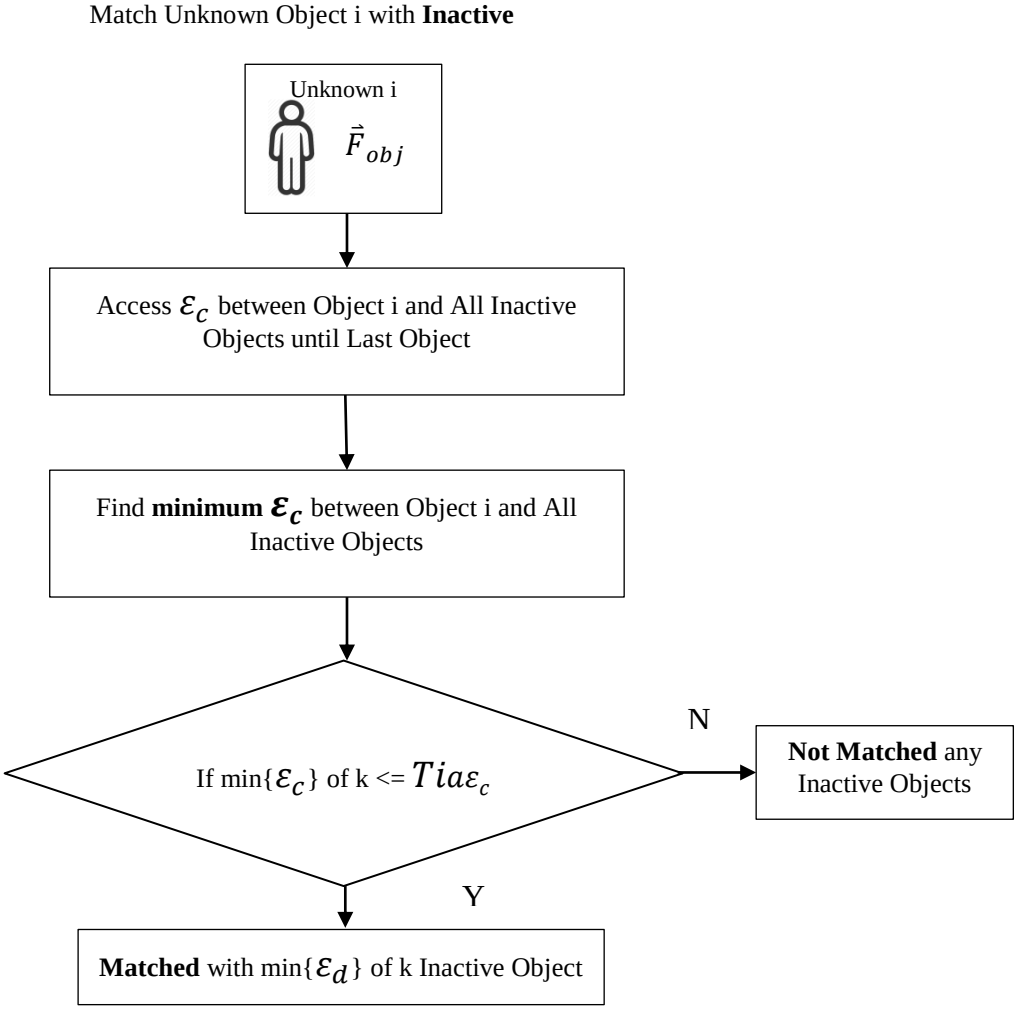
ในการติดตาม Active Object นั้นจำเป็นที่ต้องใช้ทั้ง ค่า Distance Error (ε_d) และ Color Error (ε_c) โดย Distance Error (ε_d) ช่วงกลาง-กว้างเพราะ Object มีการเคลื่อนที่ในเฟรม และ Color Error (ε_c) ช่วงกว้าง เพื่อยืนยันความถูกต้องว่าเป็นคนเดียวกัน โดยขั้นแรกระบบจะดึงค่า Distance Error และ Color Error ระหว่าง Object i กับทุก Active Object ที่คำนวณไว้แล้วในระบบ และนำมาหาค่า Minimum ของ Color Error ก็จะได้ค่าน้อยที่สุด min{ε_c} of k ที่เป็น Active Object จากนั้นใช้เกณฑ์เพื่อเปรียบเทียบในการตัดสินใจว่าวัตถุนั้นควรเป็นวัตถุ k หรือไม่ ตามค่าขีดแบ่ง Distance Threshold of Active ($Ta\varepsilon_d$) และ Color Threshold of Active ($Ta\varepsilon_c$) ซึ่งแสดงตามภาพประกอบที่ 3-26



ภาพประกอบที่ 3-26 วิธีการ Match Unknown Object i และ Active Objects

(6) การ Match Unknown Object i กับ Inactive Objects

ในการ Match Inactive Object นั้นจำเป็นที่ต้องใช้ Color Error (ε_c) เนื่องจาก Inactive เป็นวัตถุที่หายออกจากระบบไปแล้ว ไม่ได้อยู่ในเฟรมภาพ ซึ่ง Color Error (ε_c) จะมีช่วงแคบมากเพื่อยืนยันว่า Inactive ที่หายออกไปกลับมาจริง ๆ โดยขั้นแรกระบบจะดึงค่า Color Error ระหว่าง Object i กับทุก Inactive Object ที่คำนวณไว้แล้วในระบบ และนำมาหาค่า Minimum ของ Color Error ก็จะได้ค่าน้อยที่สุด min{ε_c} of k ที่เป็น Inactive Object จากนั้นใช้เกณฑ์เพื่อเปรียบเทียบในการตัดสินใจว่าวัตถุนั้นควรเป็นวัตถุ k หรือไม่ ตามค่าขีดแบ่ง Color Threshold of Active ($Tia\varepsilon_c$) ซึ่งแสดงตามภาพประกอบที่ 3-27

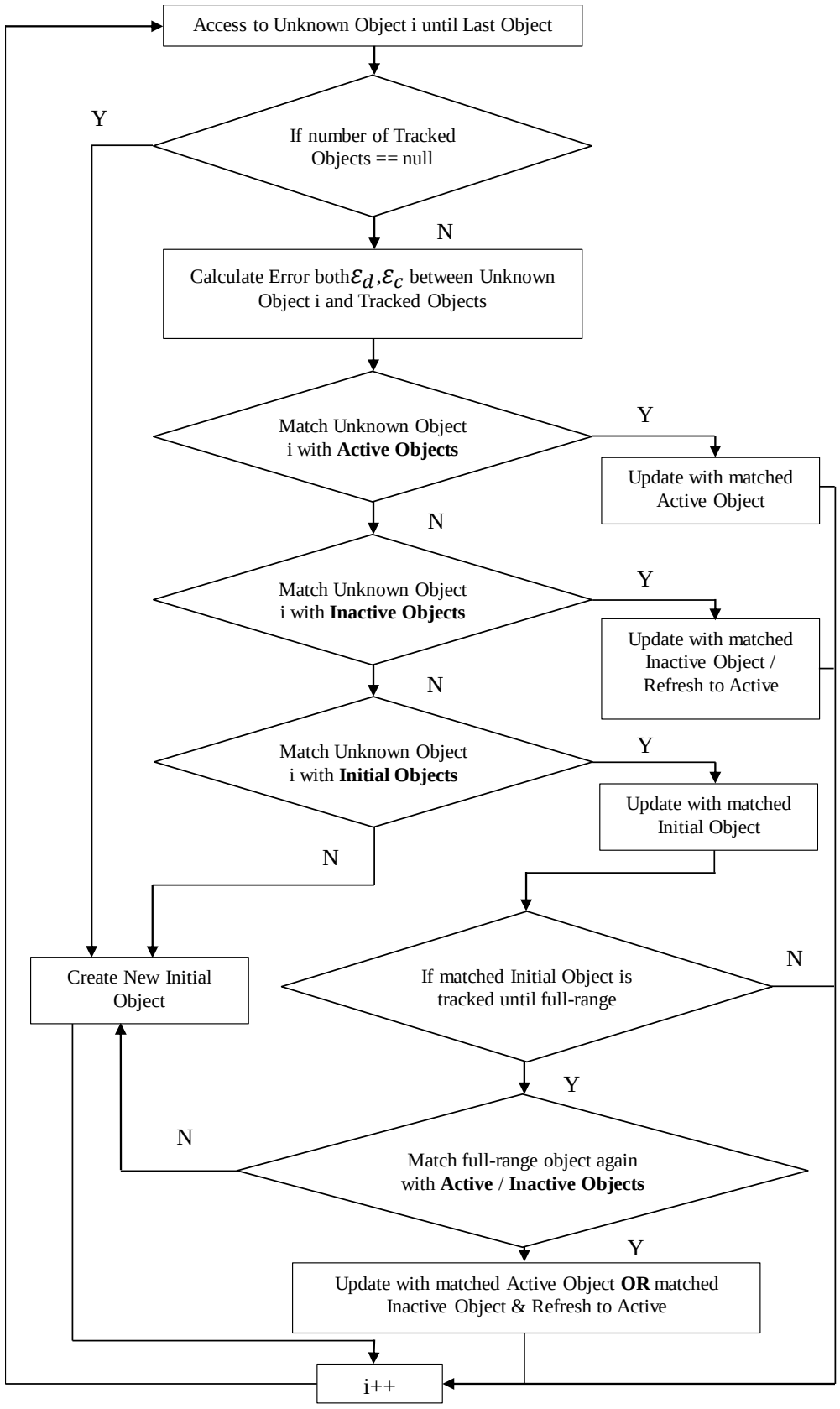


ภาพประกอบที่ 3-27 วิธีการ Match Unknown Object i และ Inactive Objects

(ง) ภาพรวมประมวลผลติดตามและจดจำตัวบุคคลในมุมมองเดี่ยว

โดยภาพรวมของการประมวลผลของติดตามและจดจำตัวบุคคลในมุมมองเดี่ยวนั้นจะเริ่มต้นที่เข้าไปที่ Unknown Object i ทีละตัวจนถึงสุดท้ายที่ตรวจจับได้ในมุมนั้น ขั้นแรกจะตรวจสอบได้ว่าเคยมีวัตถุที่ถูกตรวจจับในระบบหรือไม่ หากไม่มีจะส่งไปให้ New Object ใหม่เป็น

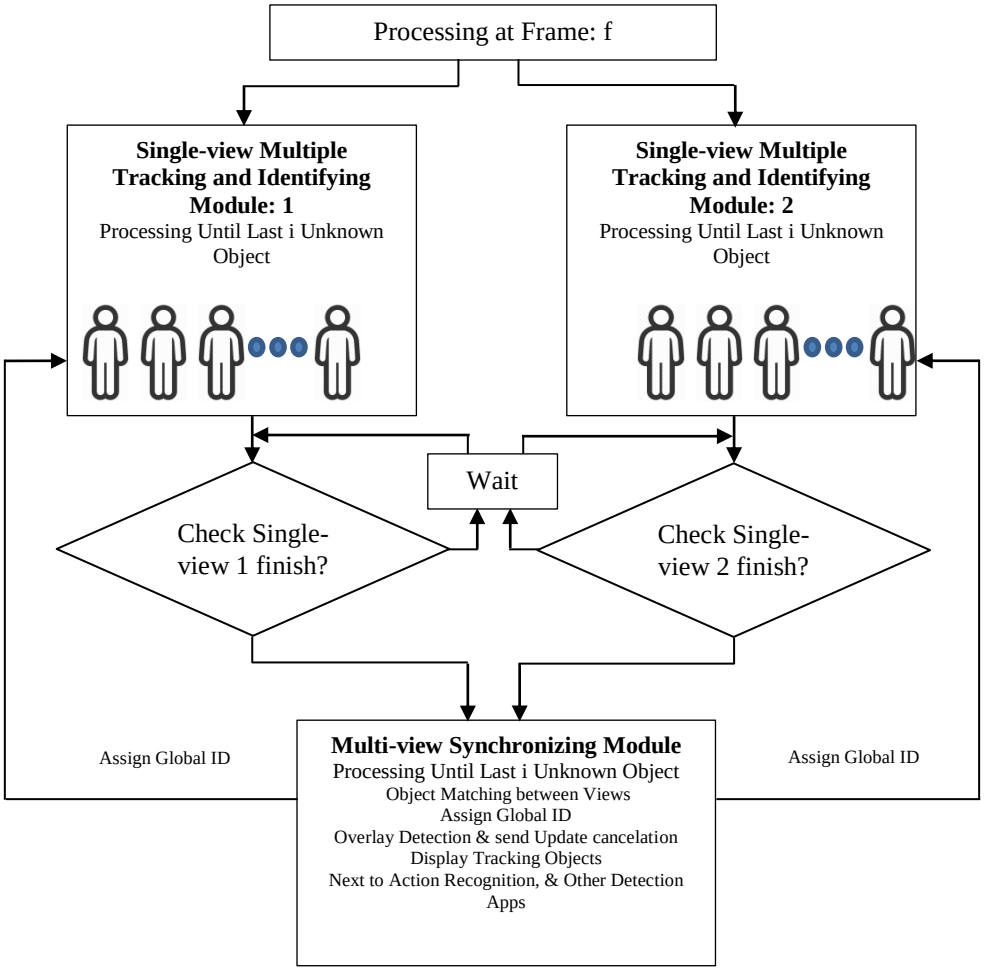
Initial หากมีอยู่ก็จะมาคำนวณค่า Error ของ Object i และวัตถุที่มีอยู่ในระบบทั้งหมด จากนั้นจะเริ่มต้น Match Unknown Object i กับ Active Objects ที่มีอยู่ในระบบก่อน หากไม่ใช่ จะ Match กับ Inactive Objects หากไม่ใช่ก็จะ Match ด้วย Initial (ถ้าหากในระบบมี Object สถานะนั้น ๆ อยู่) หาก Match ตรงกันก็จะทำการอัพเดทข้อมูลกับ Object สถานะนั้น ๆ ที่ Match ตรงกัน หาก Miss Match ก็จะทำการ New Object ใหม่เป็น Initial แต่มีในกรณีของ Match ตรงกันกับ Inactive Object และทำการอัพเดทข้อมูลแล้ว ก็จะทำการตรวจสอบต่อว่าถูกอัพเดทครบตามเวลาที่กำหนดหรือไม่ หากใช่ก็จะถูกนำไปเปรียบเทียบอีกครั้งกับ Inactive/Active object หากยังไม่ใช่ก็จะถือว่าเป็นบุคคลใหม่ซึ่งไม่เคยเข้ามาในระบบ ซึ่งได้แสดงภาพรวมการประมวลผลติดตามและจดจำตัวบุคคลในมุมมองเดี่ยว ดังอธิบายได้ตามภาพประกอบที่ 3-28



ภาพประกอบที่ 3-28 ภาพรวมการประมวลผลติดตามและจดจำตัวบุคคลในมุมมองเดี่ยว

3.3.2 ส่วนการเชื่อมโยงข้อมูลระหว่างมุมมองกล้อง (Multi-view Synchronizing Module)

ส่วนการเชื่อมโยงข้อมูลระหว่างมุมมองกล้องจะที่ทำหน้าที่เพื่อจับคู่บุคคลระหว่างกล้องและและดึงข้อมูลที่ต้องใช้ออกมา เช่น ข้อมูลที่ต้องนำไปฟิวชันกันระหว่างสองมุมเพื่อรู้จำท่าทางและข้อมูลการติดตามบุคคลในระบบ และสุดท้ายคือการตรวจจับการเกิดการบดบัง ซึ่งส่วนการเชื่อมโยงข้อมูลจะต้องรอให้การประมวลผลติดตามและจดจำตัวบุคคลในมุมมองเดี่ยวเสร็จสิ้นทุกมุมมองก่อนจึงจะเข้าถึงในส่วนนี้ได้ และประมวลผลในส่วนของส่วนการเชื่อมโยงข้อมูลระหว่างมุมมองกล้อง เนื่องจากหากระบบของการประมวลผลติดตามและจดจำตัวบุคคลในมุมมองเดี่ยวทำงานไม่เสร็จสิ้น ข้อมูลที่เชื่อมโยงกันอาจจะไม่ถูกต้องหรือไม่สมบูรณ์ตามที่ควรจะเป็น ดังแสดงกระบวนการในภาพประกอบที่ 3-29



ภาพประกอบที่ 3-29 ภาพรวมของส่วนการเชื่อมโยงข้อมูลระหว่างมุมมองกล้อง

(ก) ส่วนจับคู่บุคคลระหว่างมุมมอง (Object Matching between Views)

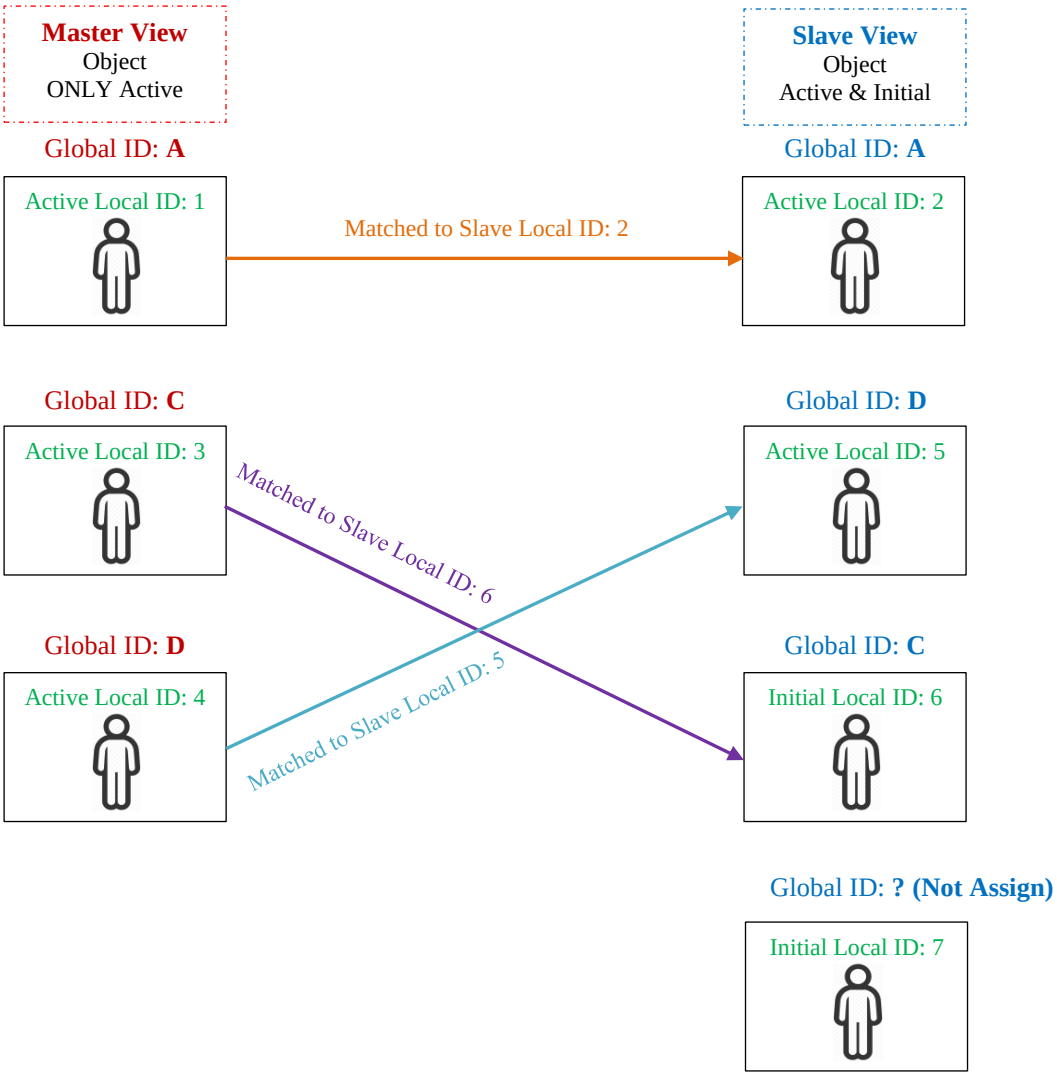
สำหรับการจับคู่ระหว่างมุมมองนั้นจะต้องมีการตั้งค่ามุมมองหลัก (Master View) เพื่อเป็นตัวตั้งในการจับคู่กับบุคคลในมุมกล้องอื่น ๆ โดย Master View จะถูกตั้งค่าตั้งแต่แรกเริ่มก่อน

โปรแกรมติดตามบุคคลจะทำงานโดยจะเลือกมุมที่มีความแม่นยำในการแยกแยะบุคคลได้ดี หรือตั้งค่าตามสภาพแสงในกล้องที่มีความเท่ากัน หรือคุณภาพของกล้องที่ดี เพื่อจะใช้เป็น Master View ซึ่งมุมมองกล้องอื่นๆจะเรียกว่า Slave View

โดยในขั้นตอนแรก Master View จะทำการ Assign Global ID ให้ Active Object ที่แปลงจาก Local ID ที่เป็นตัวเลขจำนวนเต็มบวก เป็น ตัวอักษรภาษาอังกฤษตัวใหญ่ เช่น Local ID Active = 1 Global ID = A, Local ID Active = 2 Global ID = B หลังจากนั้นจะทำการเข้าถึง Active Object ของ Master View ทีละตัวเพื่อทำการจับคู่ระหว่าง Active Object ของ Master view กับ Active / Initial Objects ของ Slave View โดยใช้ข้อมูลของคุณสมบัติของตัวบ่งชี้สี (Color Descriptor) ที่บันทึกครั้งล่าสุด ซึ่งได้แก่ $\mathrm{m}_r, \mathrm{m}_g, \mathrm{m}_b, \rho_r, \rho_g, \rho_b$ เพื่อที่จะคำนวณ Color Error ($\mathcal{E}_c$) ระหว่าง Active Object ของ Master View กับ Active / Initial Objects ของ Slave View หลังจากนั้นก็จะทำการจับคู่และแจกจ่าย Global ID ในคู่ที่ Match ตรงกันให้กับ Active / Initial Objects ของ Slave View ซึ่งแสดงตัวอย่างการจับคู่ระหว่างมุมมองตามภาพประกอบที่ 3-30

(ข) การตรวจจับการบดบัง เพื่อยกเลิกการอัพเดทข้อมูล (Overlay Detection)

เนื่องจากการบดบังหรือซ้อนกันจะทำให้ระบบทำงานผิดพลาด โดยเกิดจากวัตถุบางส่วนถูกบดบังทำให้อาจจะอัพเดทข้อมูลของบุคคลอื่น แต่เนื่องจากมุมมองของการตั้งกล้องของระบบจะเป็น 60-120 องศาต่อกันโดยประมาณในการทดสอบติดตามตัวบุคคล เพราะฉะนั้นทำให้มีโอกาสทำให้มีการบดบังน้อย และหากเกิดการบดบังมักจะมีจำนวนของวัตถุระหว่างมุมมองที่ไม่เท่ากัน ซึ่งจะใช้เป็นการตรวจจับการบดบังอย่างง่ายซึ่งประมวลผลได้รวดเร็ว โดยส่วนการเชื่อมโยงข้อมูลระหว่างมุมมองกล้องจะตรวจสอบจำนวนที่ไม่เท่ากันของวัตถุและไปทำการยกเลิกข้อมูลที่ถูกอัพเดทใน Single-View ทั้งหมด แต่อาจจะมีบางส่วนที่จะมีการบดบังเกิดขึ้นแต่มีจำนวนวัตถุระหว่างมุมมองที่เท่ากัน ซึ่งระบบการเลือกอัพเดทข้อมูลในแต่ละมุมมองในการ Matching ของ Active Object มีตัวกรองข้อมูล ซึ่งสามารถเลือกกรองข้อมูลที่มีค่า Color Error แตกต่างกันมาก ซึ่งอาจจะไม่ใช่วัตถุที่กำลังติดตามอยู่อันเนื่องมากจากการถูกบดบัง ซึ่งตัวกรองก็จะไม่อัพเดทข้อมูล



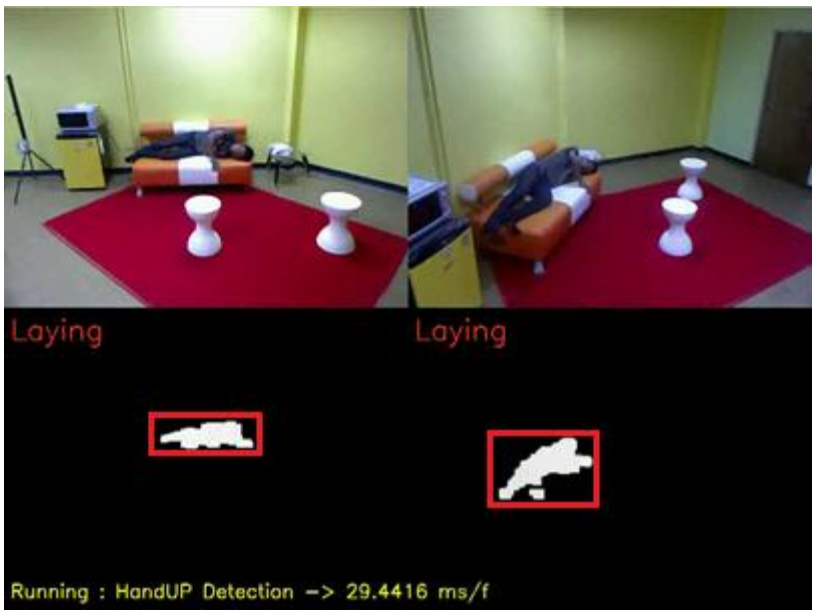
ภาพประกอบที่ 3-30 ตัวอย่างการจับคู่ระหว่างมุมมอง

3.4 การตรวจจับท่าทางที่ผิดปกติ (Abnormal Event Detection)

งานวิจัยนี้ยังมุ่งเน้นไปที่การตรวจจับเหตุการณ์ต่าง ๆ ที่มีความน่าสนใจ ซึ่งอาจจะเกิดขึ้นได้ในระบบเฝ้าระวังและดูแลด้านสุขภาพ ได้แก่ การตรวจจับการล้มซึ่งทำให้สามารถเตือนให้ผู้ที่เกี่ยวข้องสามารถให้การช่วยเหลือได้ทันท่วงที, การตรวจจับการโบกมือเพื่อขอความช่วยเหลือซึ่งในบางครั้งการผู้ต้องการความช่วยเหลืออาจจะไม่สามารถเคลื่อนที่ไปกดปุ่มที่อุปกรณ์ส่งสัญญาณขอความช่วยเหลือได้ การโบกมือให้ระบบรู้จำจากการมองเห็นจึงเป็นทางเลือกหนึ่งในการส่งสัญญาณขอความช่วยเหลือได้, และส่วนของระบบการตรวจจับการกระโดด ซึ่งอาจจะเกิดได้ในกรณีของการตกใจจากการถูกจี้ปล้น หรือวิ่งหลบบางอย่าง ซึ่งเป็นเหตุการณ์ที่ไม่ปกติสำหรับในพื้นที่ร่มในอาคาร โดยทั้งหมดจะมุ่งเน้นไปที่การนำแบบจำลองและคำตอบจากการรู้จำท่าทางพื้นฐาน ได้แก่ ยืน/ เดิน นั่ง ก้ม นอน มาประยุกต์ต่อยอดในการใช้ตรวจจับเหตุการณ์ที่ผิดปกติ โดยที่ใช้แค่ Rule-based ขั้นพื้นฐานเท่านั้น และจะทำการทดสอบในฐานข้อมูล PSU แค่เบื้องต้นเท่านั้น

3.4.1 กรณีศึกษาการล้ม (Case Study of Falling Detection)

สำหรับการตรวจจับการล้มในงานวิจัยนี้ จะใช้การประยุกต์ต่อยอดจากการวิเคราะห์ท่าทางพื้นฐาน โดยการจับลำดับการเปลี่ยนแปลงจากท่าอื่น ๆเป็นท่านอน ซึ่งต้องใช้สถานที่มาเป็นตัวยืนยันว่าการเกิดนอนหรือการล้ม โดยในงานวิจัยนี้ใช้การสร้างจุดยกเว้น ซึ่งอาจจะเป็นที่นอน โซฟา เบาะตั้งพื้นหรือเก้าอี้ โดยที่จุดยกเว้นนั้น จะใช้วิธีการให้บุคคลเข้าไปนอนในตำแหน่งที่จะยกเว้น แล้วทำการตรวจจับตำแหน่งที่ยกเว้นมาเก็บไว้ ซึ่งถ้าหากมีการนอนในบริเวณที่ยกเว้นระบบก็จะได้ไม่ตรวจจับว่าเกิดการล้มเกิดขึ้น โดยวิเคราะห์จากพื้นที่ว่าทับซ้อนในบริเวณที่ยกเว้นว่าเกินขีดเกณฑ์ที่กำหนดทั้งสองมุมหรือไม่ ดังภาพประกอบที่ 3-31 ซึ่งกรอบสี่เหลี่ยมสีแดงจะเป็นบริเวณที่ถูกยกเว้นการตรวจจับการล้ม



ภาพประกอบที่ 3-31 ตัวอย่างบริเวณที่ยกเว้นการตรวจจับการล้ม

โดยที่ในการตรวจจับการล้มจะใช้การตรวจจับการเปลี่ยนลำดับท่าทางจากท่าทางอื่น ๆที่ไม่ใช่นอนมาเป็นนอน ซึ่งในการตรวจจับว่าเกิดการล้มจริง ๆนั้น ต้องระยะรอคอยให้มั่นใจว่าได้เกิดการล้มขึ้นจริง ๆ ซึ่งสามารถตั้งค่าระยะรอคอยให้มั่นใจว่าหลังจากเปลี่ยนลำดับท่าทางจากท่าทางอื่น ๆที่ไม่ใช่ท่านอนมาเป็นท่านอนและต้องนอนติดต่อกันเท่ากับ N เฟรม จึงจะตรวจจับว่าเกิดการล้มดังตัวอย่างตามภาพประกอบที่ 3-32 ซึ่งมีค่าระยะรอคอยให้มั่นใจ N = 4 เฟรม

Other Action	Laying	Laying	Laying	Other Action	Other Action	Other Action

Falling Not Detected

Other Action	Laying	Laying	Laying	Laying	Laying	Other Action

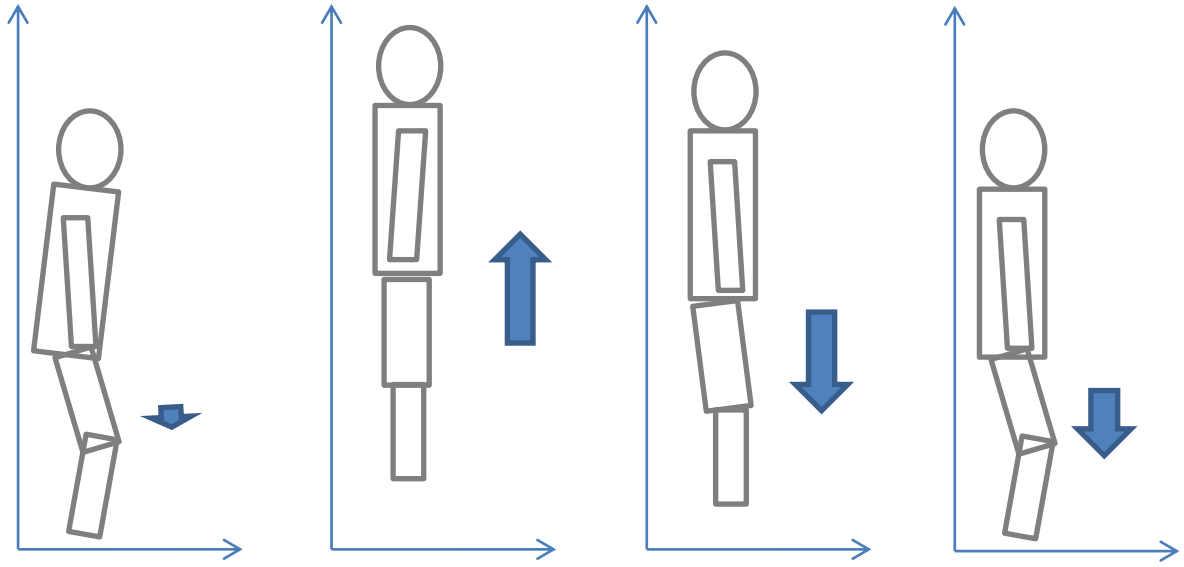
Falling Detected

โดยกำหนดให้ค่าระยะรอคอยให้มั่นใจ N = 4 เฟรม

ภาพประกอบที่ 3-32 ตัวอย่างการตรวจจับการล้มที่ตั้งค่าระยะรอคอย

3.4.2 กรณีศึกษาการกระโดด (Case Study of Jumping Detection)

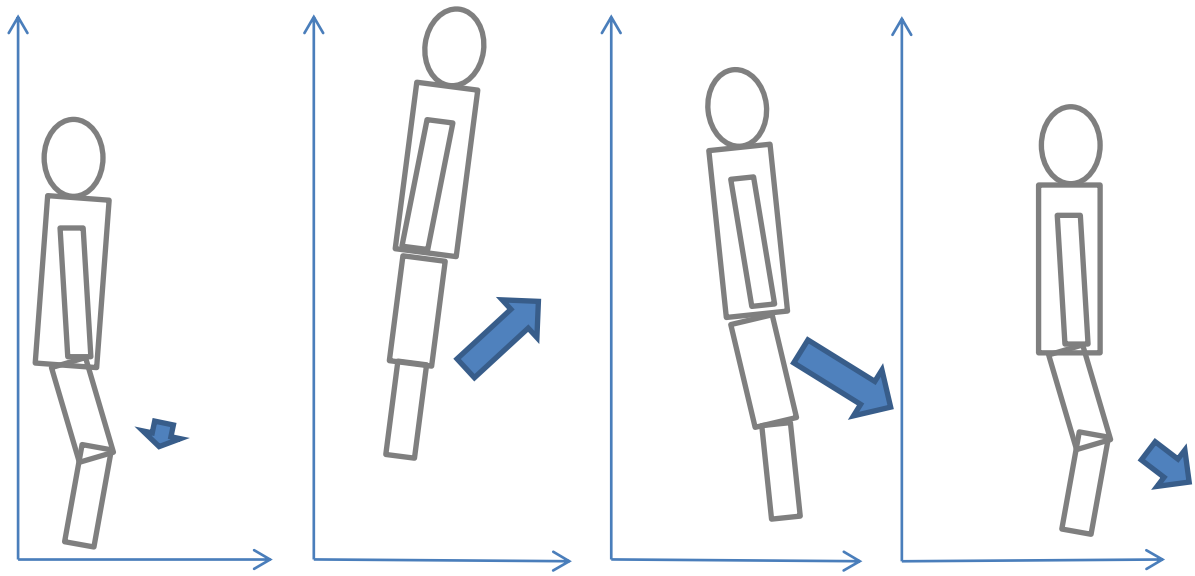
สำหรับการกระโดดจะเป็นท่าทางที่ประกอบด้วยท่าทางที่มีพื้นฐานมาจากท่ายืนเป็นหลัก ในงานวิจัยนี้จึงได้ใช้พื้นฐานของการรู้จำท่าทางพื้นฐานของมนุษย์โดยการฟิวชันฟีเจอร์ในระดับล่างจากหลายมุมมอง ในการตรวจจับการกระโดดเพื่อความถูกต้องและแม่นยำ โดยสามารถแบ่งการกระโดดได้ในขั้นต้นตามตำแหน่งได้สองลักษณะ คือกระโดดลงที่เดิม และกระโดดลงอีกที่ซึ่งมีลักษณะทิศทางแกน x และ y ที่ต่างกัน แต่สามารถใช้การวิเคราะห์ที่เหมือนกันได้ แสดงตัวอย่างดังภาพประกอบที่ 3-33 ซึ่งเป็นการกระโดดแบบลงที่เดิมซึ่งจะมีการเปลี่ยนแปลงตามแกน y เป็นส่วนใหญ่ ที่จะแตกต่างกันกับการลงอีกที่ดังภาพประกอบที่ 3-34 ซึ่งจะมีการเปลี่ยนแปลงตามแกน x และ y ที่ใกล้เคียงกันทำให้ทิศทางการพุ่งตามแกน y น้อยลง มีทิศทางที่ไปด้านหน้าด้านข้าง หรือด้านหลังก็ได้



(ก) (ข) (ค) (ง)

ภาพประกอบที่ 3-33 การกระโดดแบบลงที่เดิม

(ก) จังหวะย่อ (ข) จังหวะพุ่งขึ้น (ค) จังหวะลง (ง) จังหวะเตะพื้น



(ก) (ข) (ค) (ง)

ภาพประกอบที่ 3-34 การกระโดดแบบลงอีกที่

(ก) จังหวะย่อ (ข) จังหวะพุ่งขึ้น (ค) จังหวะลง (ง) จังหวะเตะพื้น

โดยการกระโดดจะเป็นกิจกรรมการเคลื่อนไหวในแนวดิ่งเป็นส่วนใหญ่ และใช้ช่วงระยะเวลาหนึ่งซึ่งไม่เท่ากัน ทำให้ไม่สามารถตรวจจับได้ภายในการวิเคราะห์ภาพแค่หนึ่งเฟรมการกระโดดจึงต้องตรวจจับโดยใช้จังหวะต่าง ๆ ของการกระโดดและการประเมินลักษณะจากคุณสมบัติเพิ่มเพื่อตรวจจับว่าเป็นกิจกรรมการกระโดดหรือไม่ โดยงานวิจัยนี้จะแบ่งจังหวะของการกระโดดเป็น 5 จังหวะ

1) **จังหวะย่อเพื่อดีดตัว (Knees Bending)** คือการย่อเข่าเพื่อให้มีแรงกระโดดมากขึ้น ซึ่งจะทำให้กระโดดได้สูงมากยิ่งขึ้น ซึ่งจังหวะนี้อาจจะมีการเกิดขึ้นก็ได้ หรือไม่มีการเกิดขึ้นก็ได้ แสดงให้เห็นตามภาพประกอบที่ 3.33 (ก) และ 3.34 (ก)
2) **จังหวะดีดตัวขึ้น (Dashing Up)** คือจังหวะที่มีการพุ่งตัวขึ้นไปลอยในอากาศ โดยใช้การดีดตัวขึ้นจากแรงของขาท่อนบน-ล่าง รวมถึงฝ่าเท้า เพื่อต้านแรงโน้มถ่วงและพุ่งขึ้นไปในอากาศ ซึ่งจังหวะนี้จะเกิดขึ้นเป็นเวลานานที่สุด รวมไปถึงมี vector การเคลื่อนตัวที่เห็นได้ชัดเจน แสดงตามภาพประกอบที่ 3-33 (ข) และ 3-34 (ข)
3) **จังหวะลอยตัว (Floating in Air)** คือจังหวะที่มีการลอยตัวกลางอากาศและอยู่นิ่งกับที่ ซึ่งจะเป็นเป็นช่วงเวลาสั้น ๆ อันเนื่องจากมาจากแรงยกตัวขึ้นมีค่าเท่ากันกับแรง

โน้มถ่วงที่ดึงมวลของบุคคลลงมายังพื้นโลก ซึ่งในบางครั้งไม่สามารถตรวจจับจังหวะนี้จากกล้องได้พอดีอันเนื่องมากจากความเร็วของกล้องและระยะการแบ่งและดึงเฟรมภาพ (Sampling)

4) **จังหวะดิ่งลง (Landing)** คือจังหวะที่บุคคลดิ่งลงสู่พื้นด้วยแรงโน้มถ่วงและแรงพุ่งตามแนวราบซึ่งจะเกิดขึ้นนานน้อยกว่าจังหวะดีดตัว และมี vector การเคลื่อนตัวที่เห็นได้ชัดเจน แสดงตามภาพประกอบที่ 3-33 (ค) และ 3-34 (ค)

5) **จังหวะกระทบพื้น (Hit on the Ground)** เป็นจังหวะที่เท้าลงถึงพื้น โดยอาจจะมีการย่อเข่าเล็กน้อยเพื่อรับแรงกระแทกที่อาจจะเกิดขึ้นก็ได้ตามภาพประกอบที่ 3-33 (ง) และ 3-34 (ง)

เมื่อพิจารณาจากจังหวะทั้ง 5 แล้วสามารถใช้จังหวะที่เป็นองค์ประกอบของการกระโดดที่มีการเกิดขึ้นอย่างชัดเจน เพื่อตรวจจับว่ามีการกระโดด ซึ่งมี 2 จังหวะคือ จังหวะดีดตัวขึ้น (Dashing Up) และจังหวะดิ่งลง (Landing) ที่มีการเกิดขึ้นทุกครั้งที่มีการกระโดดและสามารถตรวจจับได้โดยกล้องสีและความลึก โดยถ้าทั้งสองเหตุการณ์นี้ต้องเกิดขึ้นตามลำดับคือเกิดจังหวะดีดตัวขึ้น (↑) และมีระยะห่างที่รอจังหวะดิ่งลง (*) ตามกำหนด ก็จะตรวจจับว่ามีการกระโดดเกิดขึ้น แต่หากเกิดจังหวะดีดตัวขึ้น (↑) และไม่มีจังหวะดิ่งลง (↓) หรือเกินระยะห่างของเฟรมที่อยู่ในช่วงที่กำหนด ก็จะมีการล้างจังหวะดีดตัวขึ้น (-) และรอจังหวะดีดตัวขึ้น (Dashing Up) ต่อไป ซึ่งสามารถอธิบายได้ตามภาพประกอบที่ 3-35

-	-	-	-	↑	*	*	↓	-	-

Jump Detected

-	-	-	-	↑	*	*	*	↓	-

Jump Detected

-	-	-	↑	*	*	*	-	↓	-

Jump Not Detected

-	-	↑	*	*	*	-	-	↑	*

Jump Not Detected

* wait interval threshold <= 3

ภาพประกอบที่ 3-35 ตัวอย่างการตรวจจับการกระโดดจากการจับจังหวะ

ในงานวิจัยนี้จะแบ่งระบบการตรวจจับการกระโดดเป็น 3 ขั้นตอนหลัก ๆ ดังนี้

1) **การรับข้อมูลท่าทางพื้นฐานและข้อมูลบุคคลจากการรู้จำท่าทาง (Information Acquisition from Action Recognition)** ที่เป็นข้อมูลพื้นฐานต่าง ๆของบุคคลตรวจจับได้ในขั้นตอนของการรู้จำท่าทางโดยการฟิวชันฟีเจอร์ในระดับล่างจากหลายมุมมอง ในหัวข้อที่ 3.2 เช่น ท่าทาง, ตำแหน่งของบุคคล, Motion Depth Object เป็นต้น

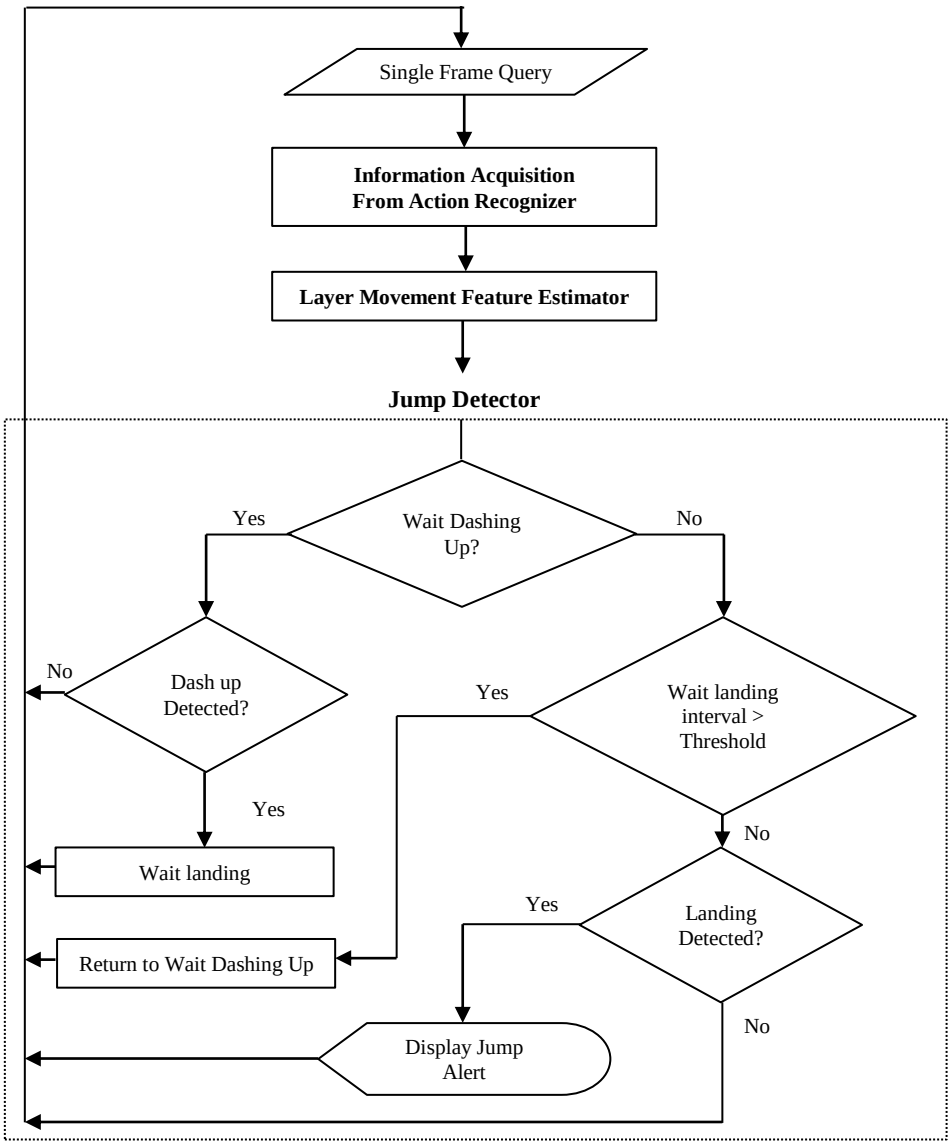
2) **การดึงคุณลักษณะการเคลื่อนไหวของวัตถุตาม Layer (Layer Movement Feature Estimator)** เป็นการสร้างแบบจำลองที่ติดตามการเคลื่อนไหวของวัตถุ โดยมีคุณลักษณะ เช่น จุดศูนย์กลางแกน (Axis), ตำแหน่งในแต่ละเฟรม (Position), ระยะทางที่เคลื่อนที่ (Distance), ความเร็ว (Velocity) เป็นต้น

3) **การตรวจจับการกระโดด (Jump Detection)** เป็นการตรวจจับกิจกรรมการกระโดดโดยใช้การตรวจจับจังหวะดีดตัวขึ้น (Dashing Up) และจังหวะดิ่งลง (Landing)

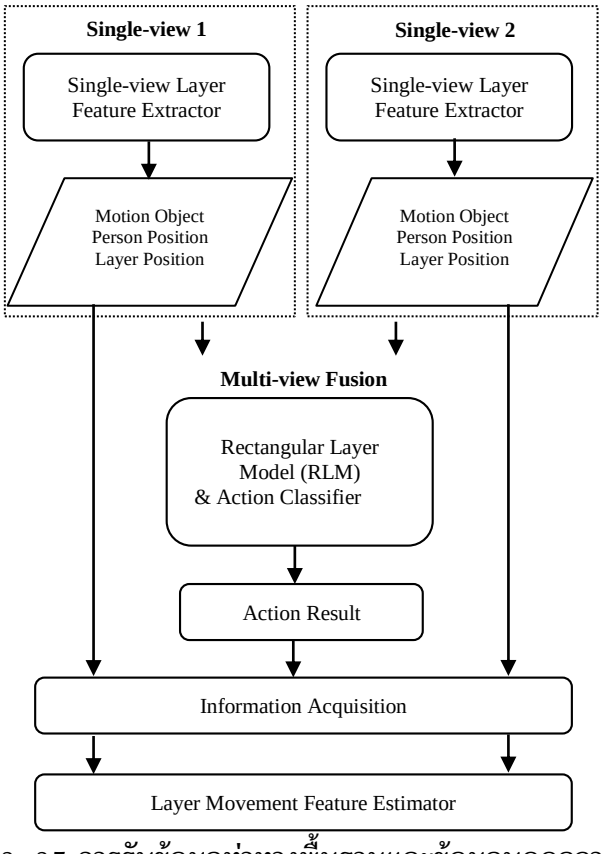
โดยขั้นตอนหลัก ๆเหล่านี้สามารถอธิบายได้ตามแผนผังงานการตรวจจับการกระโดด (Jump Detection Flowchart) ที่แสดงไว้ในภาพประกอบที่ 3-36

(ก) การรับข้อมูลท่าทางพื้นฐานและข้อมูลบุคคลจากการรู้จำท่าทาง (Information Acquisition from Action Recognition)

ในขั้นตอนนี้เป็นการเชื่อมต่อกับการรู้จำท่าทางพื้นฐานของมนุษย์ เพื่อนำข้อมูลที่มีอยู่มาใช้ในการสร้างแบบจำลองการเคลื่อนไหวของวัตถุโดยจะนำข้อมูลที่สำคัญที่ได้จากการตรวจจับบุคคล (Reduced-Noise Motion Object) ซึ่งได้จากการตรวจจับการเคลื่อนไหวและได้รับการลดทอนสัญญาณรบกวน รวมไปถึงนำข้อมูลตำแหน่งของ Layer ต่าง ๆ ตามสมการที่ (3.19) และ (3.20) ในหัวข้อที่ 3.2 การรู้จำท่าทางโดยการฟิวชันฟีเจอร์ในระดับล่างจากหลายมุมมอง มาหาค่าฟีเจอร์ Axis ซึ่งเป็นตำแหน่งตรงกลางของเลเยอร์ ที่แสดงไว้ตามสมการที่ (3.54) และ (3.55) และผลลัพธ์ที่ได้จากการรู้จำท่าทาง (Action Result) โดยทั้งสองมุมมองเหล่านี้จะถูกดึงมาใช้เพื่อการสร้างแบบจำลองที่ติดตามการเคลื่อนไหวของวัตถุ ซึ่งอธิบายได้ตามภาพประกอบที่ 3-37



ภาพประกอบที่ 3-36 แผนผังงานการตรวจจับการกระโดด



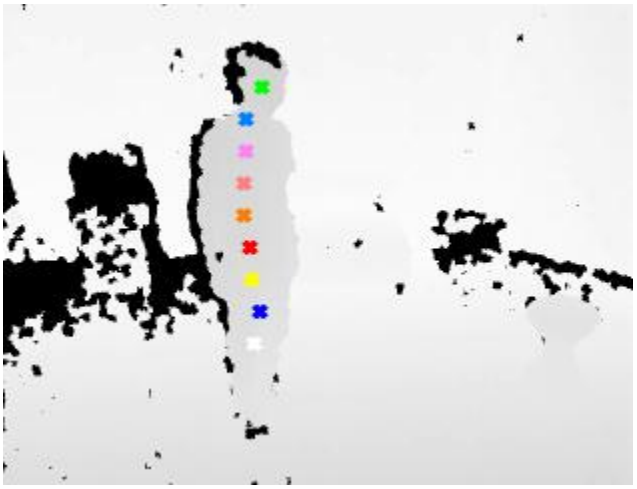
ภาพประกอบที่ 3-37 การรับข้อมูลท่าทางพื้นฐานและข้อมูลบุคคลจากการรู้จำท่าทาง

(ข) การดึงคุณลักษณะการเคลื่อนไหวของวัตถุตาม Layer (Layer Movement Feature Estimator)

ส่วนนี้จะดึงคุณสมบัติในแต่ละมุมมองเพื่อสร้างแบบจำลองที่ติดตามการเคลื่อนไหวของวัตถุ โดยในขั้นแรกจะต้องหาจุดศูนย์กลางแกน (Axis) ซึ่งใช้เป็นตัวแทนตำแหน่งที่อยู่ของแต่ละ Layer ในเฟรมหนึ่ง ๆโดยจะอยู่ในตำแหน่งตรงกลางของวัตถุใน Layer นั้น ๆ ซึ่งเป็นการหาค่าจุดพิกเซลที่มีความสว่าง ($\mathcal{B}$) ด้านซ้าย ($l_{\mathcal{B}}$) และค่าจุดพิกเซลที่มีความสว่างด้านขวา ($r_{\mathcal{B}}$) เพื่อจะหาค่าจุดศูนย์กลางที่เป็นค่าจุดศูนย์กลางแกน ($C_{\perp}$) ดังสมการที่ (3.54) และ (3.55) ซึ่งตัวอย่างจุดศูนย์กลางแกน (Axis) ในแต่ละ Layer แสดงดังภาพประกอบที่ 3-38

$$C_{\perp}x(i) = x + \frac{r_{\mathcal{B}} - l_{\mathcal{B}}}{2} \tag{3.54}$$

$$C_{\perp}y(i) = {}_{y_T[\,k\,]+}\frac{y_B[\,k\,] + y_T[\,k\,]}{2} \tag{3.55}$$



ภาพประกอบที่ 3-38 ตัวอย่างจุดศูนย์กลางแกน (Axis) ในแต่ละ Layer

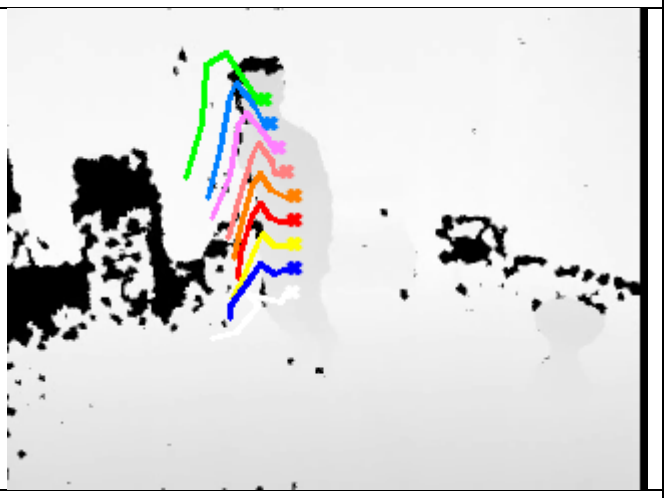
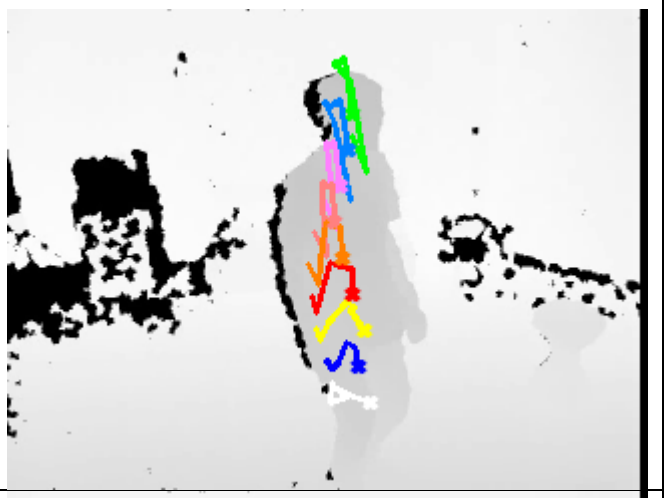
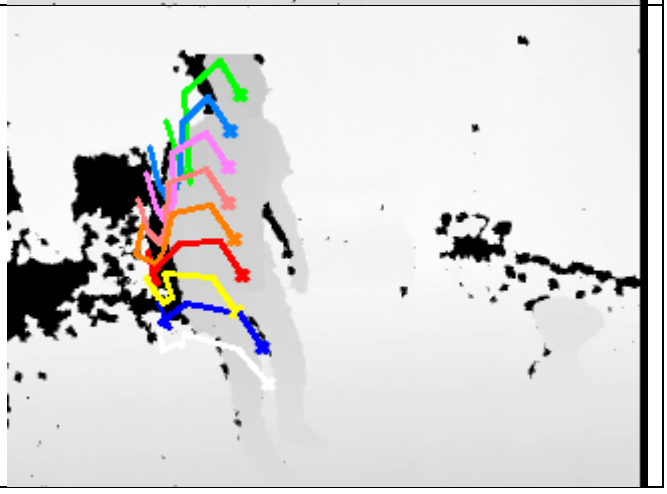
สามารถหาจุดศูนย์กลางแกน (Axis) ในแต่ละ Layer ได้แล้วทุก ๆเฟรมก็จะมีตำแหน่งที่เป็นตัวแทนในแต่ละ Layer ไว้ หลังจากนั้นให้หาค่าระยะทางที่เคลื่อนที่ใน Layer ระหว่างเฟรมที่ f และ $f-1$ ดังสมการที่ (3.56) และ (3.57)

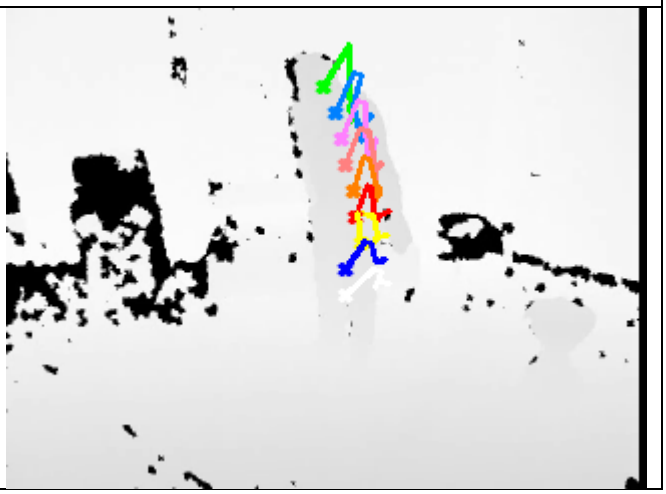
$$S_x(f) = C_\perp x(f) - C_\perp x(f-1) \tag{3.56}$$

$$S_y(f) = C_\perp y(f) - C_\perp y(f-1) \tag{3.57}$$

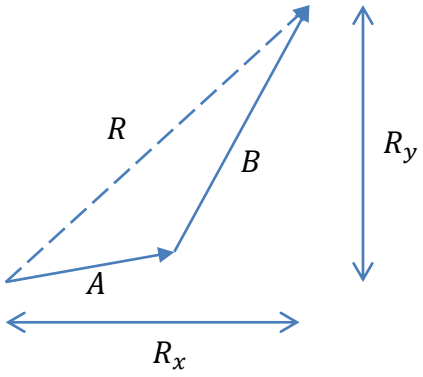
ซึ่งเมื่อพิจารณารูปแบบจากระยะทาง (Distance) ช่วงเวลาหนึ่ง ๆแล้วจะมีรูปแบบของระยะทางที่เป็น Vector ที่สามารถพิจารณาว่าเป็นรูปแบบของการกระโดด ซึ่งแสดงตัวอย่างดังตารางที่ 3-8

ตารางที่ 3-8 ตัวอย่างรูปแบบ Vector ของการกระโดด

รูปแบบการกระโดด	ตัวอย่างภาพประกอบ
กระโดดลงอีกที่ไปด้านหลัง	
กระโดดลงที่เดิมไปด้านหน้า	
กระโดดลงอีกที่ไปด้านหน้า	

รูปแบบการกระโดด	ตัวอย่างภาพประกอบ
กระโดดลงที่เดิมไปด้านหน้า	

หลังจากที่หาระยะทางที่เคลื่อนที่ใน Layer ระหว่างเฟรมที่ f และ$f-1$ $(S_{xy}(f))$ ได้แล้ว ขั้นตอนต่อมาคือการรวมระยะทางเหล่านี้โดยใช้การรวมแบบเวกเตอร์ ซึ่งจะทำให้ได้ Vector ที่แสดงคุณลักษณะของการเคลื่อนที่ซึ่งจะมีความสัมพันธ์ในแนวแกนของทั้งสองมุมมอง โดยแทนระยะทางจากทั้งสองมุมเป็น Vector A และ B ดังภาพประกอบที่ 3-39 และหาค่าตามสมการที่ (3.58) และ (3.59)



ภาพประกอบที่ 3-39 ภาพแทนระยะทางจากทั้งสองมุมเป็น Vector A และ B

$$R_x = A_x + B_x \tag{3.58}$$

$$R_y = A_y + B_y \tag{3.59}$$

เวกเตอร์ย่อย ๆในแต่ละ Layer $R_x(i)$ และ $R_y(i)$ สามารถรวมเข้าด้วยกันเพื่อหา Vector รวมที่บ่งบอกทิศทาง และขนาดโดยรวม ซึ่งจะนำไปใช้ในการพิจารณาจังหวะของการกระโดด

ต่าง ๆต่อไปได้ สำหรับการรวม Vector ในแต่ละ Layer จะเป็นการบวกกันในแต่ละแนวแกนตามสมการที่ (3.60) และ (3.61)

$$Ve_x = \sum_{i=0}^{N} R_x(\mathrm{i}) \tag{3.60}$$

$$Ve_y = \sum_{i=0}^{N} R_y(\mathrm{i}) \tag{3.61}$$

เมื่อเวกเตอร์ย่อย ๆในแต่ละ Layer ถูกรวมเป็น Ve_x และ Ve_y ซึ่งเป็น Vector โดยรวม และจะหาขนาด (Magnitude) และทิศทาง (Direction) ได้ตามสมการที่ (3.62) และ (3.63) ดังแสดงตัวอย่าง Vector โดยรวม ได้ตามภาพประกอบที่ 3-40

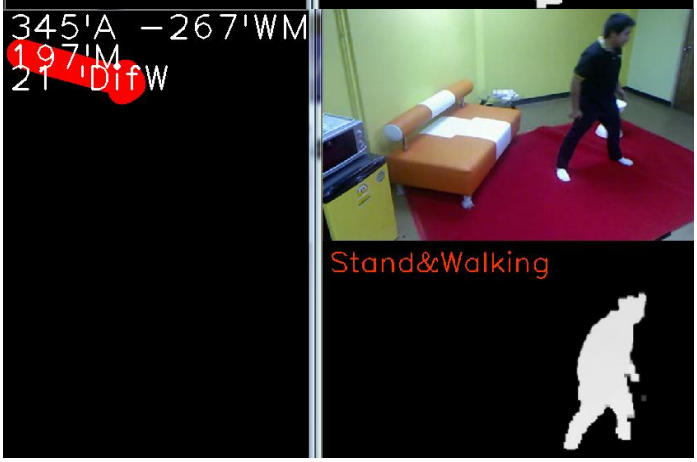
$$|Ve| = \sqrt{{Ve_x}^2 + {Ve_y}^2} \tag{3.62}$$

$$\theta_{Ve} = \tan^{-1}\frac{Ve_y}{Ve_x} \tag{3.63}$$

(ก)

(ข)

(ค)

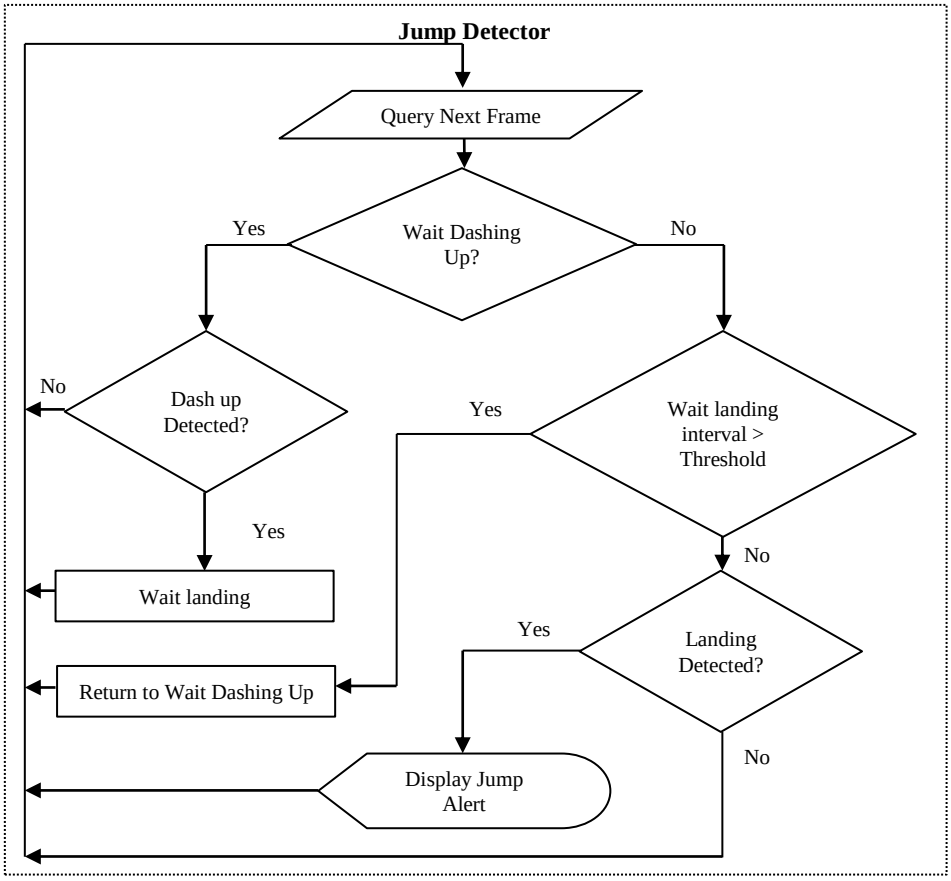


ภาพประกอบที่ 3-40 ตัวอย่าง Vector โดยรวมของการกระโดด

(ก) จังหวะดีดตัวขึ้น (ข) จังหวะดิ่งลง (ค) การเดิน

(ค) การตรวจจับการกระโดด (Jump Detection)

การตรวจจับการกระโดดในงานวิจัยนี้จะเป็นการตรวจจับกิจกรรมการกระโดดโดยใช้การตรวจจับจังหวะดีดตัวขึ้น (Dashing Up) และ จังหวะดิ่งลง (Landing) มีกฎว่าต้องเกิด 2 จังหวะนี้จะต้องต่อเนื่องกันตามลำดับจึงจะตรวจจับว่าเป็นการกระโดด ซึ่งสามารถแสดงได้ดังภาพประกอบที่ 3-41



ภาพประกอบที่ 3-41 ผังงานของขั้นตอนการตรวจจับการกระโดด

สำหรับการตรวจจับการดีดตัวขึ้น และจังหวะดิ่งลง จะมีการคำนวณค่าต่าง ๆที่ได้จากแบบจำลองคุณลักษณะการเคลื่อนไหวของวัตถุตาม Layer ที่รวมเวกเตอร์ย่อย ๆในแต่ละ Layer ให้เป็น Ve_x และ Ve_y ซึ่งเป็น Vector โดยรวม โดยสามารถหาขนาด (Magnitude) และทิศทาง (Direction) ได้ นอกจากนี้ยังใช้คุณสมบัติโดยรวมของข้อมูลตำแหน่งของบุคคล (Person Position) ตามหัวข้อที่ 3.2.1 ซึ่งเป็นพิกัดของกรอบสี่เหลี่ยม (Bounding Rectangle) ที่ประกอบไปด้วยตำแหน่งบนซ้าย X_h, Y_h, ความกว้าง W_h, ความสูง H_h , และนำผลลัพธ์ที่ได้จากการรู้จำท่าทาง (Action Result) มาใช้ในการช่วยให้การพิจารณาการตรวจจับกระโดดมีความแม่นยำ ซึ่งมี

คุณสมบัติต่าง ๆที่ใช้เป็น เกณฑ์พิจารณาจังหวะดีดตัวขึ้น / ดิ่งลง และใช้ในการถ่วงน้ำหนักค่าความเชื่อมั่นเพื่อตรวจจับว่าเป็นจังหวะดีดตัวขึ้น / ดิ่งลง ดังที่จะแสดงในตารางที่ 3-9

ตารางที่ 3-9 คุณสมบัติต่าง ๆที่ใช้เป็นเกณฑ์พิจารณาและถ่วงน้ำหนักค่าความเชื่อมั่นจังหวะดีดตัวขึ้น / ดิ่งลง

คุณสมบัติต่าง ๆที่ใช้ถ่วงน้ำหนักค่าความเชื่อมั่น	**คุณสมบัติต่าง ๆที่ใช้เกณฑ์พิจารณา**
ทิศทาง (Direction) ของ Vector โดยรวม (θ_{Ve})	
ความเร็วถ่วงน้ำหนัก (υ_w)	จำนวนของการรู้จำเป็นท่ายืนในเวลานั้น (n_{st})
การเปลี่ยนแปลงตำแหน่งของขาในแนวดิ่ง (Db_y)	-

ซึ่งค่าความเร็วถ่วงน้ำหนักเป็นการใช้ระยะทางจากเฟรมที่แล้วครึ่งหนึ่ง และระยะทางจากเฟรมปัจจุบัน โดยได้จากค่าขนาด (Magnitude) ของ Vector โดยรวม ($|Ve|$) แสดงได้ตามสมการที่ (3.64)

$$\upsilon_w = |Ve|(f) + 0.5|Ve|(f-1) \tag{3.64}$$

ค่าจำนวนของการรู้จำเป็นท่ายืนในเวลานั้น (n_{st}) จะได้จากผลลัพธ์ของการรู้จำที่ออกมาในช่วงเวลาก่อนหน้าถึงเฟรมปัจจุบันที่ออกมาเป็นท่ายืน ซึ่งโดยพื้นฐานของการกระโดดต้องมาจากท่ายืน จึงจำเป็นต้องใช้คุณสมบัตินี้เป็นเกณฑ์ในการพิจารณา ซึ่งแสดงตามสมการที่ (3.65)

$$n_{st} = \sum_{fi=fn-f}^{fi} Action(fi)_{\stackrel{\text{def}}{=} stand} \tag{3.65}$$

โดยที่ fi คือ เฟรมที่กำลังดำเนินการหาค่า

f คือ เฟรมปัจจุบัน

fn คือ เฟรมก่อนหน้าที่ย้อนไปจำนวน n

โดยค่าการเปลี่ยนแปลงตำแหน่งของขาในแนวดิ่ง (Db_y) จะเป็นค่าที่ได้จากตำแหน่งจากคุณสมบัติโดยรวมของข้อมูลตำแหน่งของบุคคล (Person Position) ซึ่งจะระบุเป็นกรอบของวัตถุ

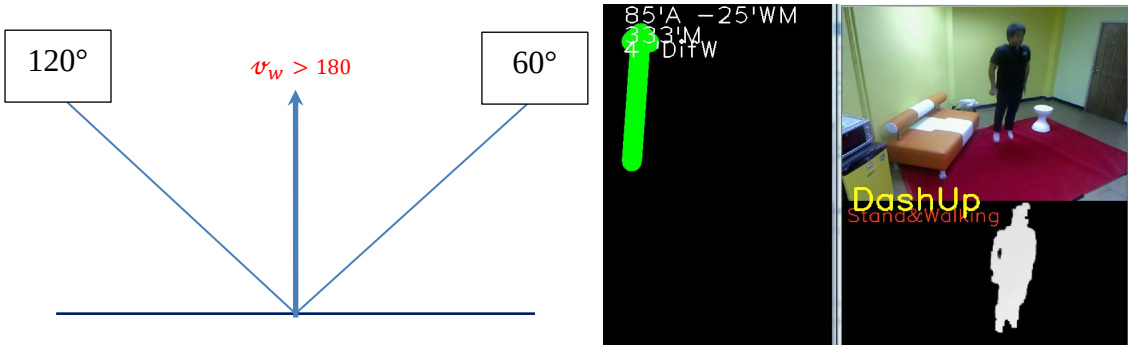
(Bounding Box) ซึ่งจะมีตำแหน่งของขาที่เป็นตำแหน่งล่างสุดของวัตถุได้ในแต่ละเฟรม (b_y) และคำนวณหาค่าความเปลี่ยนแปลงเหล่านี้จากเฟรมภาพก่อนหน้า (Db_y) ดังสมการที่ (3.66)

$$Db_y = \sum_{fi=f}^{f-fn} |b_y(fi) - b_y(fi-1)| \tag{3.66}$$

ซึ่งคุณสมบัติที่ได้กล่าวมาข้างต้นจะใช้ในการเป็นกฎเกณฑ์ และใช้ถ่วงน้ำหนักค่าความเชื่อมั่น ในการตรวจจับจังหวะดีดตัวขึ้น และจังหวะดิ่งลง ซึ่งเป็นค่าที่ใช้ในการตัดสินใจว่าเป็นจังหวะจังหวะดีดตัวขึ้น และจังหวะดิ่งลงหรือไม่ ที่จะได้กล่าวในส่วนถัดไป

(1) การตรวจจับจังหวะดีดตัวขึ้น (Dashing Up)

เป็นจังหวะที่มีการพุ่งตัวขึ้นไปลอยในอากาศ ซึ่งจังหวะนี้จะเกิดขึ้นนานที่สุด รวมไปถึงมี vector การเคลื่อนตัวที่เห็นได้ชัดเจน โดยจากการทดลองและพิจารณาของทิศทาง (Direction) ของ Vector โดยรวม (θ_{Ve}) จะอยู่ระหว่าง 60°–120° และในเบื้องต้นควรจะมีความเร็วถ่วงน้ำหนัก (v_w) มากกว่า 180 ซึ่งจะใช้เป็นกฎเกณฑ์เพื่อคัดกรองในขั้นต้นก่อนที่จะถ่วงน้ำหนักค่าความเชื่อมั่น เพื่อตัดสินใจว่าเป็นจังหวะจังหวะดีดตัวขึ้น หรือไม่ ซึ่งอธิบายได้ตามภาพประกอบที่ 3-42



ภาพประกอบที่ 3-42 กฎเกณฑ์เพื่อคัดกรองในขั้นต้นก่อนที่จะถ่วงน้ำหนักค่าความเชื่อมั่นจังหวะที่มีการพุ่งตัวขึ้น

หลังจากผ่านกฎเกณฑ์เพื่อคัดกรองในขั้นต้นแล้วจะต้องมีการถ่วงน้ำหนักค่าความเชื่อมั่นโดยใช้ค่าทิศทาง (Direction) ของ Vector โดยรวม (θ_{Ve}), ความเร็วถ่วงน้ำหนัก (v_w) และการเปลี่ยนแปลงตำแหน่งของขาในแนวดิ่ง (Db_y) โดยจะได้ค่าถ่วงน้ำหนักค่าความเชื่อมั่นของจังหวะดีดตัวขึ้น (ϖ_D) ที่มาจากสมการที่ (3.67)

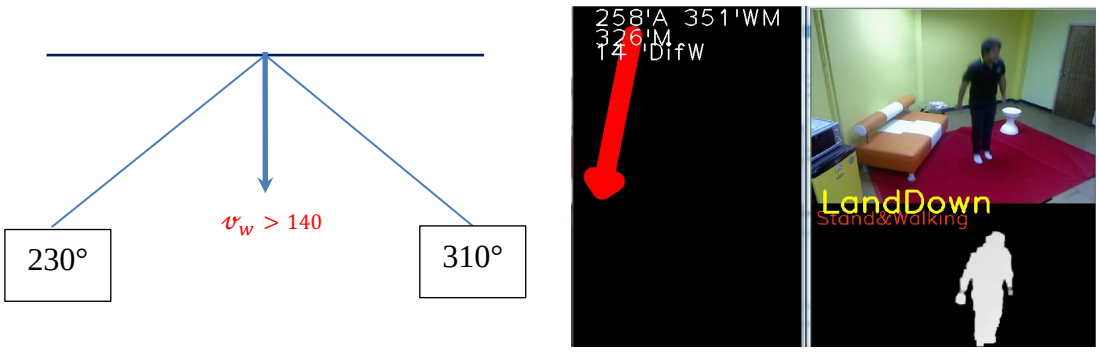
$$\varpi_D = \upsilon_w - (\,3|\theta_{\mathrm{Ve}} - 90|\,) + Db_y \qquad (3.67)$$

เมื่อได้ค่าถ่วงน้ำหนักค่าความเชื่อมั่นของจังหวะดีดตัวขึ้น (ϖ_D) แล้วจะพิจารณาว่ามีค่าถึงเกณฑ์ที่กำหนดว่ามีค่าถึง 210 (ค่าที่ได้จากการทดลอง) และพิจารณาร่วมกับจำนวนของการรู้จำเป็นท่ายืนในเวลานั้น (n_{st}) ว่ามีค่าถึง 90% (ค่าที่ได้จากการทดลอง) หากผ่านเกณฑ์ทั้งสองค่านี้แล้วก็จะถือว่าได้ตรวจจับเจอจังหวะจังหวะดีดตัวขึ้น (Dashing Up) อธิบายตามสมการที่ (3.68)

$$f(\varpi_D) = \begin{cases} \varpi_D \geq 210 \cup n_{st} \geq 0.9 \rightarrow Detected \\ elsewhere \rightarrow Not\ Detected \end{cases} \qquad (3.68)$$

(2) การตรวจจับจังหวะดิ่งลง (Landing)

จังหวะดิ่งลงเป็นจังหวะที่บุคคลดิ่งลงสู่พื้นด้วยแรงโน้มถ่วงและแรงพุ่งตามแนวราบซึ่งจะเกิดขึ้นนานน้อยกว่าจังหวะดีดตัว และมี Vector การเคลื่อนตัวที่เห็นได้ชัดเจน โดยจากการทดลองและพิจารณาของทิศทาง (Direction) ของ Vector โดยรวม (θ_{Ve}) จะอยู่ระหว่าง 230°-310° และในเบื้องต้นควรจะมีความเร็วถ่วงน้ำหนัก (υ_w) มากกว่า 140 ซึ่งจะใช้เป็นกฎเกณฑ์เพื่อคัดกรองในขั้นต้นก่อนที่จะถ่วงน้ำหนักค่าความเชื่อมั่น เพื่อตัดสินใจว่าเป็นจังหวะจังหวะดิ่งลงหรือไม่ ซึ่งอธิบายได้ตามภาพประกอบที่ 3-43



ภาพประกอบที่ 3-43 กฎเกณฑ์เพื่อคัดกรองในขั้นต้นก่อนที่จะถ่วงน้ำหนักค่าความเชื่อมั่นจังหวะที่มีการดิ่งลง

หลังจากผ่านกฎเกณฑ์เพื่อคัดกรองในขั้นต้นแล้วจะต้องมีการถ่วงน้ำหนักค่าความเชื่อมั่นโดยใช้ค่าทิศทาง (Direction) ของ Vector โดยรวม (θ_{Ve}), ความเร็วถ่วงน้ำหนัก (υ_w) และการเปลี่ยนแปลงตำแหน่งของขาในแนวดิ่ง (Db_y) โดยจะได้ค่าถ่วงน้ำหนักค่าความเชื่อมั่นของจังหวะดิ่งลง (ϖ_L) ที่มาจากสมการที่ (3.69)

$$\varpi_L = \nu_w - (\,3|\theta_{\mathrm{Ve}} - 270|\,) + \, Db_y \tag{3.69}$$

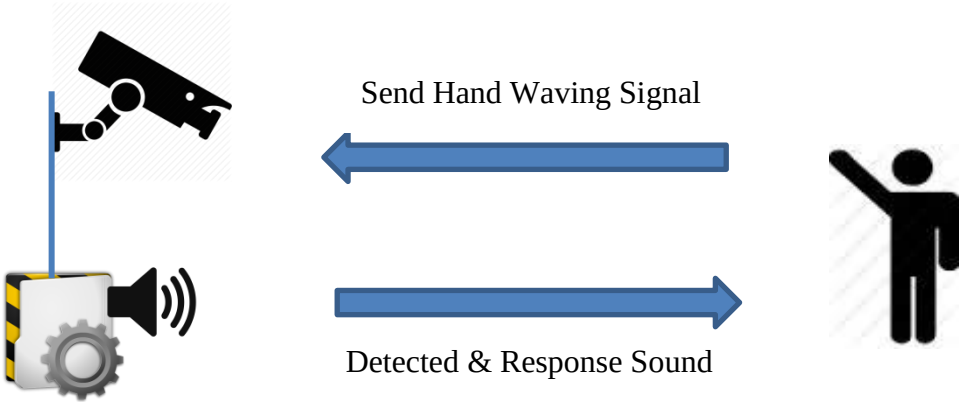
เมื่อได้ค่าถ่วงน้ำหนักค่าความเชื่อมั่นของจังหวะดิ่งลง (ϖ_L) แล้วจะพิจารณาว่ามีค่าถึงเกณฑ์ที่กำหนดว่ามีค่าถึง 170 (ค่าที่ได้จากการทดลอง) และพิจารณาร่วมกับจำนวนของการรู้จำเป็นท่ายืนในเวลานั้น (n_{st}) ว่ามีค่าถึง 90% (ค่าที่ได้จากการทดลอง) หากผ่านเกณฑ์ทั้งสองค่านี้แล้วก็จะถือว่าได้ตรวจจับเจอจังหวะดิ่งลง (Landing) อธิบายตามสมการที่ (3.70)

$$f(\varpi_L) = \begin{cases} \varpi_L \geq 170 \cup \; n_{st} \geq 0.9 \; \rightarrow Detected \\ elsewhere \rightarrow Not\ Detected \end{cases} \tag{3.70}$$

3.4.3 กรณีศึกษาการโบกมือขอความช่วยเหลือ (Case Study of Hand Waving Detection)

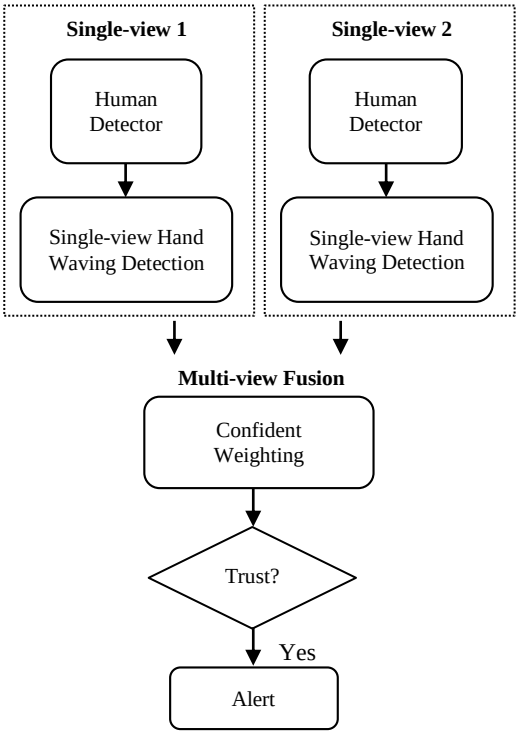
การตรวจจับการโบกมือเป็นท่าทางที่ใช้แขนยกขึ้นเหนือศีรษะคล้ายกับยกมือแล้วโบกไปมา ซึ่งสามารถใช้ในการขอความช่วยเหลือได้ โดยการโบกมือนี้จะทำในท่าทางพื้นฐานใด ๆก็ได้ ไม่ว่าจะเป็นท่า ยืน / เดิน นั่ง นอน และก้ม แต่ถ้าเป็นท่าก้มจะการโบกมือได้ยากลำบากกว่าท่าพื้นฐานอื่น ๆเนื่องจากสรีระของร่างกายในการก้มไม่เอื้ออำนวยต่อการยกแขนเหนือศีรษะ โดยในงานวิจัยนี้จะไม่ใช้การติดตามและระบุในส่วนต่าง ๆของร่างกาย โดยเฉพาะแขน เพราะต้องการใช้ความเร็วในการตรวจจับ แต่จะใช้การวิเคราะห์ลักษณะเชิงพื้นที่ของบุคคลเพื่อตรวจจับว่าบริเวณที่น่าจะเป็นแขนอยู่ในบริเวณใกล้เคียงกับศีรษะหรือไม่ โดยสามารถตรวจจับเจอภายในเฟรม ๆเดียวและไม่ต้องใช้คุณสมบัติในแต่ละมุมมองเพื่อสร้างแบบจำลองที่ติดตามการเคลื่อนไหวของวัตถุ

เนื่องด้วยระบบจำเป็นที่จะต้องมีความจำเพาะ (Specificity) สูง ที่จะต้องไม่เตือนขึ้นมาหากไม่มีการโบกจริง (False Positive) และต้องมีการตอบรับไปยังผู้ที่ขอความช่วยเหลือว่าระบบรับรู้การขอความช่วยเหลือ ดังนั้นในงานวิจัยนี้จึงจำเป็นต้องมีระบบตอบรับด้วยเสียงหากผู้ใช้โบกมือขอความช่วยเหลือ เพื่อที่ให้สามารถปรับแต่งระบบให้มีความจำเพาะ (Specificity) สูง กล่าวคือผู้ที่โบกมือขอความช่วยเหลือจะต้องพยายามโบกมือจนกว่าตอบรับ และผู้โบกมือขอความช่วยเหลือก็จะรับรู้ได้ว่าระบบตรวจจับได้แล้วตามภาพประกอบที่ 3-44



ภาพประกอบที่ 3-44 การทำงานของระบบการตรวจจับการโบกมือขอความช่วยเหลือ

สำหรับการตรวจจับการโบกมือนั้นมีข้อจำกัดในเรื่องมุมมอง ซึ่งในบางมุมมองไม่สามารถเห็นแขนที่ยกขึ้นมาโบกได้อย่างชัดเจนเนื่องจากมีการบดบัง ดังนั้นระบบจึงต้องใช้ข้อมูลจากหลายมุมมองเข้ามาช่วยการวิเคราะห์ โดยแต่ละมุมมองจะทำงานแยกส่วนไม่เกี่ยวข้องกันเพื่อตรวจจับการโบกมือ และจะนำผลลัพธ์ของการตรวจจับจากเฟรมปัจจุบันและอดีตมาวิเคราะห์ค่าความเชื่อมั่นเพื่อยืนยันว่ามีการโบกมือจริง ๆ ซึ่งแสดงการดังผังงานตามภาพประกอบที่ 3-45



ภาพประกอบที่ 3-45 ผังงานแสดงการแบ่งส่วนการทำงานเพื่อตรวจจับการโบกมือ

3.5.1 ตรวจจับการโบกมือในมุมมองเดี่ยว (Single-view Hand Waving Detection)

ในการตรวจจับการโบกมือในมุมมองเดี่ยวจะต้องวิเคราะห์ โดยใช้ข้อมูลที่ได้จากรู้จำท่าทางพื้นฐานของมนุษย์ โดยจะนำข้อมูลที่สำคัญที่ได้จากการตรวจจับบุคคล ซึ่งได้จากการตรวจจับการเคลื่อนไหวและได้รับการลดทอนสัญญาณรบกวนในหัวข้อที่ 3.2 มาใช้เพื่อใช้ในการวิเคราะห์ลักษณะเชิงพื้นที่ของบุคคลเพื่อตรวจจับว่าบริเวณที่น่าจะเป็นแขนอยู่ในบริเวณใกล้เคียงกับศีรษะมากน้อยเพียงใด

โดยการวิเคราะห์จะเริ่มจากการวิเคราะห์การวางแนวของบุคคล (Placement Direction) ว่าบุคคลตั้งอยู่ในแนวตั้ง (Vertical) หรือแนวราบ (Horizontal) เนื่องจากตำแหน่งของแขนในแนวตั้งและแนวนอนจะอยู่ในตำแหน่งที่ต่างกัน ซึ่งทำให้ต้องมีการวิเคราะห์ที่ต่างกันออกไปตามการวางแนว โดยการวิเคราะห์การวางแนวของบุคคล จะใช้ข้อมูลอัตราส่วนระหว่าง ความกว้างและความสูง ($\propto$) ซึ่งแสดงดังสมการที่ (3.71)

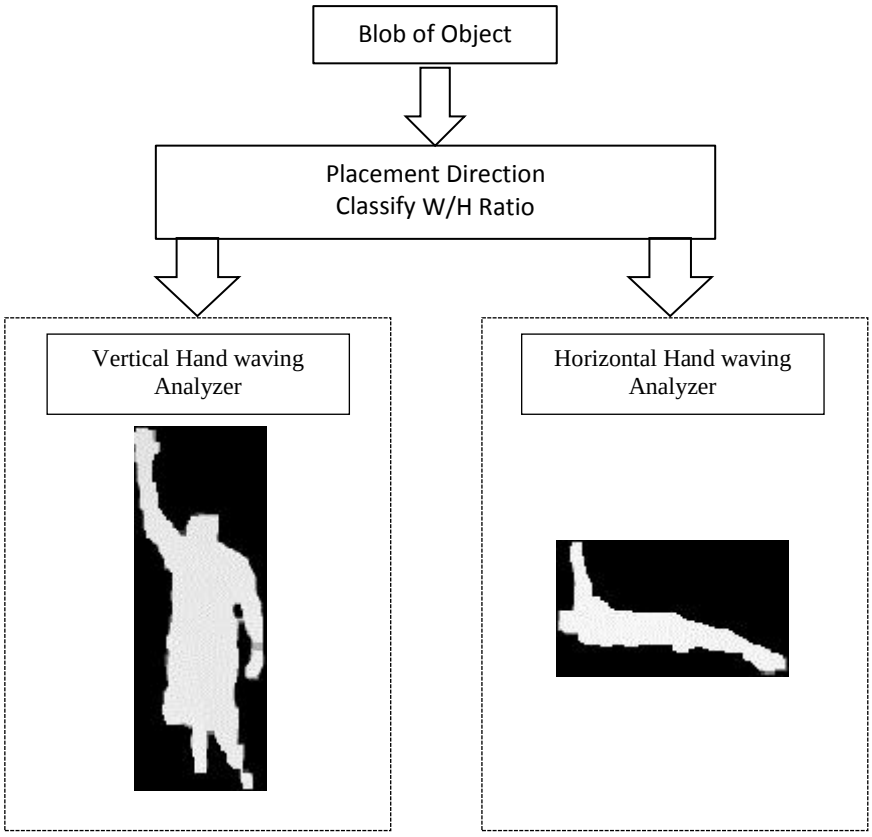
$$\propto = \frac{c}{r} \tag{3.71}$$

โดยที่ c คือ จำนวนหลักของ Object

r คือ จำนวนแถวของ Object

หลังจากที่ได้ค่าอัตราส่วนระหว่าง ความกว้างและความสูง ($\propto$) แล้วจะวิเคราะห์โดยใช้การตั้งค่าเกณฑ์ที่กำหนด (Thresholding) เพื่อระบุว่าการวางแนวของบุคคลเป็นแนวตั้ง (Vertical) หรือแนวราบ (Horizontal) ตามสมการที่ (3.72) ซึ่งอธิบายได้ตามภาพประกอบที่ 3-46

$$\propto = \begin{cases} \propto > 1.3, & Horizontal \\ \propto \leq 1.3, & Vertical - \end{cases} \tag{3.72}$$



ภาพประกอบที่ 3-46 การวิเคราะห์การโบกมือตามการวางแนวของบุคคล

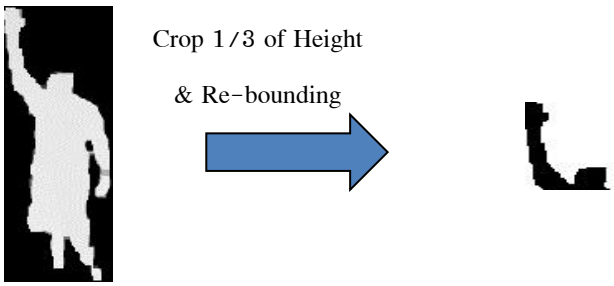
(ก) การวิเคราะห์การโบกมือตามแนวตั้ง (Vertical Hand Waving Analysis)

ในขั้นตอนนี้จะเป็นการวิเคราะห์ลักษณะเชิงพื้นที่ของบุคคล โดยถ้าหากมีการยกมือเพื่อโบกมือบริเวณที่เป็นแขนอยู่ในบริเวณใกล้เคียงกับศีรษะ ซึ่งถ้ามีการโบกมือที่มีการวางแนวของบุคคลตามแนวตั้ง จะทำให้บริเวณส่วนบนของ Bounding Box of Object มีสัดส่วนที่เป็นตัวบุคคลน้อย เพราะจะมีแต่บริเวณที่เป็นแขนปรากฏ

สำหรับกระบวนการตรวจจับจะใช้การวิเคราะห์โดยการตั้งค่าเกณฑ์ที่กำหนด เพื่อวิเคราะห์ว่าส่วนบนของ Bounding Box of Object เป็นส่วนที่มีแขนปรากฏอยู่หรือไม่ โดยแบ่งเป็น 4 ขั้นตอนหลัก คือ

(1) การตัดเฉพาะส่วนบนของ Bounding Box of Object

ในขั้นตอนนี้จะเป็นการตัดเฉพาะส่วนบน 1 ใน 3 ของความสูง และความกว้าง 100% ของตัว Bounding Box of Object ซึ่งจะได้บริเวณที่เป็นส่วนบนเพื่อนำไปหามีสัดส่วนที่เป็นตัวบุคคลต่อพื้นที่ต่อไป ซึ่งแสดงตัวอย่างตามภาพประกอบที่ 3-47



ภาพประกอบที่ 3-47 การตัดเฉพาะส่วนบนของ Bounding Box of Object

(2) การนำส่วนบนของ Bounding Box of Object ไป Project ในแนวตั้ง (Vertical Projection)

หลังจากได้พื้นที่เฉพาะส่วนบนแล้วก็จะนำพื้นที่นั้นมา Project ในแนวตั้งตามสมการที่ (3.73) และ Normalize ด้วยค่าความสูงของพื้นที่เฉพาะส่วนบน ตามสมการที่ (3.74) ได้อธิบายตามภาพประกอบที่ 3-48

$$W_{ver}(j) = \sum_{j=1}^{n} f(I(i,j) == 1) \tag{3.73}$$

โดยที่ i คือ แถว

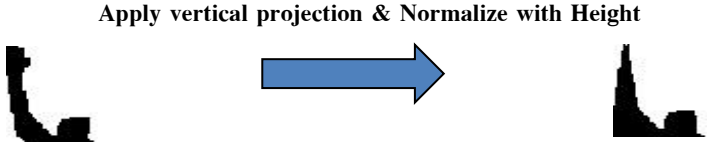
j คือ หลัก

n คือ จำนวนหลักทั้งหมด

m คือ จำนวนแถวทั้งหมด

$W_{ver}(j)$ คือ ค่าน้ำหนักในโปรเจคชั่นในหลักที่ j

$$W_{n\_ver}(j) = \frac{W_{ver}(j)}{m} \tag{3.74}$$



ภาพประกอบที่ 3-48 การนำส่วนบนของ Bounding Box of Object ไป project ในแนวตั้ง (Vertical Projection)

(3) การคำนวณหาค่าความหนาแน่นของพื้นที่ของ Bounding Box of Object จาก Vertical Projection

เป็นการ Integrate ส่วนย่อย ๆที่ถูก Decomposition ซึ่งอยู่ในรูปแบบของ Vertical Projection และหารด้วยพื้นที่ทั้งหมด เมื่อแต่ละค่าแทนด้วยความหนาแน่นในหลักหนึ่ง ๆของพิกเซล จึงสามารถรวมเพื่อหาค่าพื้นที่ของบุคคลที่มีอยู่ในส่วนบนของ Bounding Box of Object แสดงการ Integrate ได้ตามสมการที่ (3.75)

$$A_{ver} = \frac{\sum_{j=1}^{n} W_{n\_ver}(j)}{n} \tag{3.75}$$

(4) การวิเคราะห์เพื่อระบุการโบกมือตามแนวตั้ง โดยใช้การตั้งค่าเกณฑ์ที่กำหนด (Thresholding)

เป็นการวิเคราะห์ว่ามีสัดส่วนที่เป็นตัวบุคคลต่อพื้นที่มากน้อยเพียงใด หากมีน้อยกว่าค่า Threshold ซึ่งเป็นค่าที่ได้จากการทดลอง มีค่าเท่ากับ 0.35 ก็แสดงว่าบริเวณส่วนบนของ Bounding Box of Object นั้นน่าจะมีแขนที่ยกขึ้นเหนือศีรษะอยู่ แสดงดังสมการที่ (3.76)

$$A_{ver} = \begin{cases} A_{ver} < 0.35, & Detected \\ A_{ver} \geq 0.35, & ---- \end{cases} \tag{3.76}$$

(ข) การวิเคราะห์การโบกมือตามแนวราบ (Horizontal Hand Waving Analysis)

ในขั้นตอนนี้จะลักษณะการวิเคราะห์ลักษณะเชิงพื้นที่ของบุคคลคล้ายกับ การวิเคราะห์การโบกมือตามแนวตั้ง แต่จะมีขั้นตอนการหาตำแหน่งเพื่อตัดส่วนบนที่ใช้วิเคราะห์ที่ต่างกัน โดยถ้าหากมีการยกมือเพื่อโบกมือบริเวณที่เป็นแขนอยู่ในบริเวณใกล้เคียงกับศีรษะ ซึ่งถ้ามีการโบกมือที่มีการวางแนวของบุคคลตามแนวราบ จะทำให้บริเวณส่วนบนของ Bounding Box of Object มีสัดส่วนที่เป็นตัวบุคคลน้อยมาก ๆ

สำหรับกระบวนการตรวจจับจะใช้การวิเคราะห์โดยการตั้งค่าเกณฑ์ที่กำหนด เพื่อวิเคราะห์ว่าส่วนบนของ Bounding Box of Object เป็นส่วนที่มีแขนปรากฏอยู่หรือไม่เช่นเดียวกับการวิเคราะห์การโบกมือตามแนวตั้ง โดยแบ่งเป็น 4 ขั้นตอนหลัก คือ

(1) การนำ Bounding Box of Object ไป project ในแนวราบ (Horizontal Projection)

เป็นขั้นตอนที่นำพื้นที่ทั้งวัตถุมา Project ในแนวราบตามสมการที่ (3.77) ก็จะได้ Horizontal Projection ดังแสดงตัวอย่างได้ตามภาพประกอบที่ 3-49

$$W_{hor}(i) = \sum_{i=1}^{m} f(I(i,j) == 1) \quad (3.77)$$

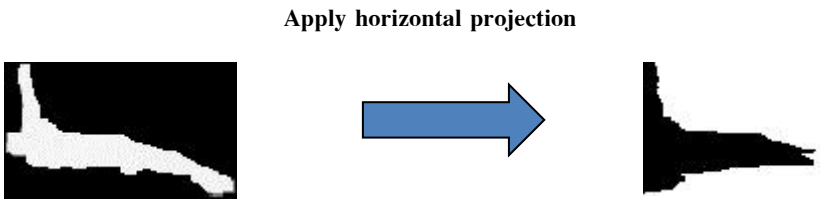
โดยที่ i คือ แถว

j คือ หลัก

n คือ จำนวนหลักทั้งหมด

m คือ จำนวนแถวทั้งหมด

$W_{hor}(i)$ คือ ค่าน้ำหนักในโปรเจคชั่นในแถวที่ i

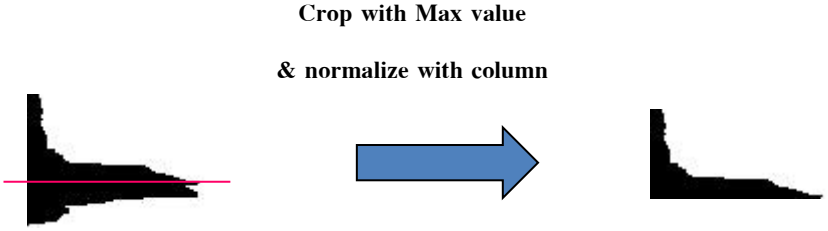


ภาพประกอบที่ 3-49 Horizontal Projection ของ Bounding Box of Object

(2) การวิเคราะห์หาส่วนบนของ Bounding Box of Object ที่เหมาะสมจากค่าของ Horizontal Projection

เนื่องจากบุคคลที่อยู่แนวราบจะมีลักษณะการวางตัวในหลายรูปแบบจึงต้องใช้การวิเคราะห์หาค่าจุดตัวที่เหมาะสมเพื่อให้ได้บริเวณส่วนบนของ Object ที่แท้จริง ในขั้นตอนนี้จะหาค่าตำแหน่งของ Horizontal Projection ที่มีค่ามากที่สุด และตัดส่วนอื่น ๆของค่าใน Horizontal Projection ออก และ Normalize ด้วยค่าความกว้างของ Bounding Box of Object ตามสมการที่ (3.78) ซึ่งจะได้บริเวณที่เป็นส่วนบนเพื่อนำไปหามีสัดส่วนที่เป็นตัวบุคคลต่อพื้นที่ต่อไป ซึ่งแสดงตัวอย่างตามภาพประกอบที่ 3-50

$$W_{n\_hor}(i) = \frac{W_{ver}(i)}{m} \quad (3.78)$$



ภาพประกอบที่ 3-50 การวิเคราะห์หาส่วนบนของ Bounding Box of Object โดยใช้ Horizontal Projection

(3) การคำนวณหาค่าความหนาแน่นของพื้นที่ของ Bounding Box of Object จาก Horizontal Projection

เป็นการ Integrate ส่วนแต่ละแถวจาก Decomposition ซึ่งอยู่ในรูปแบบของ Horizontal Projection ในบริเวณส่วนบนของ Bounding Box of Object ที่เหมาะสม และหารด้วยพื้นที่ทั้งหมด โดยแต่ละค่าของ Horizontal Projection แทนด้วยความหนาแน่นในแถวหนึ่ง ๆของพิกเซล จึงสามารถรวมเพื่อหาค่าพื้นที่ของบุคคลที่มีอยู่ในส่วนบนของ Bounding Box of Object แสดงการ Integrate ได้ตามสมการที่ (3.79)

$$A_{hor} = \frac{\sum_{i=1}^{n} W_{n\_hor}(i)}{m} \tag{3.79}$$

(4) การวิเคราะห์เพื่อระบุการโบกมือตามแนวนอน โดยใช้การตั้งค่าเกณฑ์ที่กำหนด (Thresholding)

เป็นการวิเคราะห์ว่ามีสัดส่วนที่เป็นตัวบุคคลต่อพื้นที่มากน้อยเพียงใด หากมีน้อยกว่าค่า Threshold ซึ่งเป็นค่าที่ได้จากการทดลอง มีค่าเท่ากับ 0.35 ก็แสดงว่าบริเวณส่วนบนของ Bounding Box of Object นั้นน่าจะมีแขนที่ยกขึ้นเหนือบริเวณลำตัวในแนวราบอยู่ แสดงดังสมการที่ (3.80)

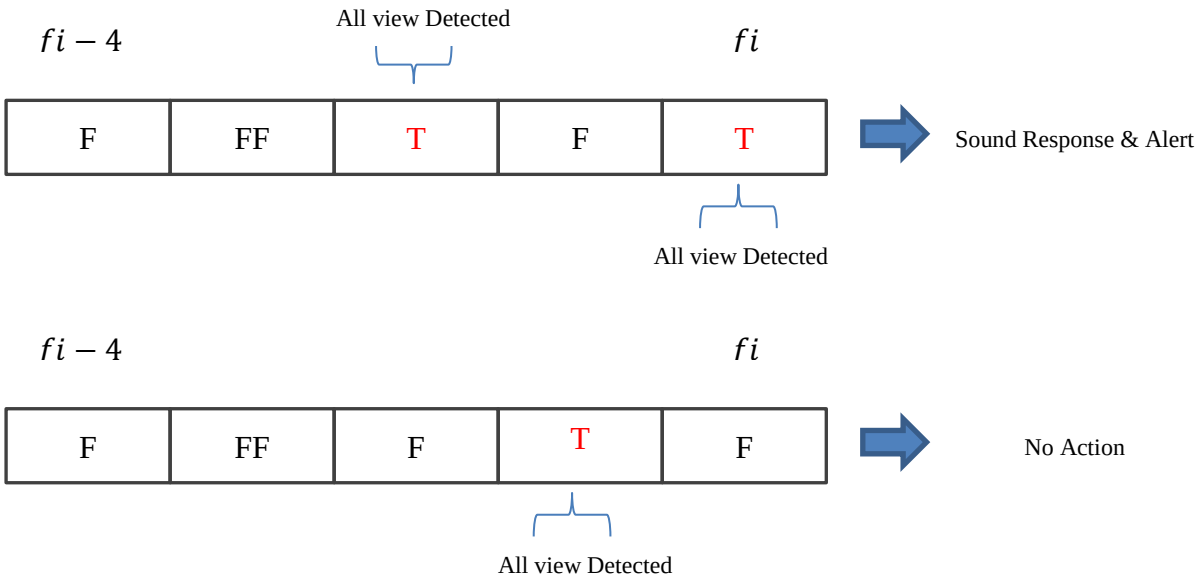
$$A_{hor} = \begin{cases} A_{hor} < 0.35, & Detected \\ A_{hor} \geq 0.35, & ---- \end{cases} \tag{3.80}$$

(ค) การตรวจจับการโบกมือโดยใช้การวิเคราะห์ค่าความมั่นใจจากหลายมุมมอง (Hand Waving Detection using Confident Weighting from Multi-view)

หลังจากที่ส่วนของระบบตรวจจับการโบกมือในมุมมองเดี่ยว (Single-view Hand Waving Detection) วิเคราะห์เพื่อตรวจจับการโบกมือแล้ว ระบบจะนำผลลัพธ์ของการตรวจจับจากทุกมุมมองเดี่ยวมาวิเคราะห์ว่าความเชื่อมั่นว่ามีการโบกมือมากน้อยเพียงใด โดยใช้ค่าผลลัพธ์ของการตรวจจับจากเฟรมปัจจุบันและอดีตมาวิเคราะห์เพื่อยืนยันว่ามีการโบกมือจริง ๆ หลังจากนั้นก็จะทำการตอบรับด้วยเสียงและแจ้งเตือนไปยังระบบอื่นเพื่อติดต่อให้ความช่วยเหลือต่อไป

โดยการยืนยันว่าเกิดการโบกมือจริงจะประกอบไปด้วย 2 องค์ประกอบคือ สามารถตรวจจับการโบกมือจากทุกมุมมองเดี่ยว และระยะเวลาที่โบกต้องนานพอ ดังนั้นเมื่อระบบตรวจจับได้ว่ามีการโบกมือในทุกมุมมองก็จะเก็บค่าไว้ว่าในเฟรมปัจจุบันเกิดการโบกมือ (T) และหากตรวจสอบจำนวนที่การโบกมือใน Sequential Result จากเฟรมปัจจุบัน fi และอดีตมาที่เก็บ

ค่าไว้จำนวน fn แล้วมีค่ามากกว่าหรือเท่ากับ Threshold (T_c) ก็จะทำการตอบรับด้วยเสียงและแจ้งเตือน โดยค่า T_c ที่เหมาะสมจากการทดลองคือมากกว่าหรือเท่ากับ 2 และfn ซึ่งเป็นจำนวน history ใน Sequential Result ที่เหมาะสมจากการทดลองมีค่าเท่ากับ 5 ซึ่งแสดงตัวอย่างการตรวจจับได้ตามภาพประกอบที่ 3-51



ภาพประกอบที่ 3-51 ตัวอย่างการตรวจจับโดยใช้การวิเคราะห์ค่าความมั่นใจจากหลายมุมมอง

3.5 สรุป

ในบทนี้ได้นำเสนอวิธีในการวิจัยการวิเคราะห์ท่าทางจากหลายมุมมอง ทั้งการปรับปรุงการวิเคราะห์ท่าทางจากระบบวิเคราะห์มุมมองเดี่ยวโดยการฟิวชันข้อมูลในระดับสูง(คำตอบ)จากหลายมุมมอง โดยจะมีแบบจำลองที่เป็นฟังก์ชันเพิ่มความถูกต้องของการรู้จำโดยสร้างฟังก์ชันวัดค่าความน่าเชื่อถือของคำตอบจากโดยการวัดหาค่ามุมมอง และใช้ข้อมูลจากการวัดความถูกต้องในการรู้จำท่าทางในแต่ละมุมมอง ดูความแม่นยำจาก Confusion Matrix อีกทั้งได้นำเสนอการรู้จำท่าทางโดยการฟิวชันฟีเจอร์ในระดับล่างจากหลายมุมมอง โดยใช้แบบจำลองการฟิวชันแบบเลเยอร์ ที่แบ่งพื้นที่เพื่อ Encode ฟีเจอร์ลงในเลเยอร์ในมุมมองเดี่ยวและนำไปฟิวชันเพื่อปรับปรุงเป็นฟีเจอร์ใหม่โดยใช้ข้อมูลจากหลายมุมมอง จากนั้นเวกเตอร์ของฟีเจอร์เหล่านี้จะถูกนำไปใช้ในขั้นตอนการเรียนรู้จดจำและแยกแยะ เพื่อสามารถจะนำมาใช้ในการรู้จำท่าทางต่อไป นอกจากนี้ผู้วิจัยยังได้นำเสนอการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง ที่เป็นการติดตามในกรณีที่มีบุคคลเข้ามาในพื้นที่มากกว่าหนึ่งคน เพื่อจับคู่บุคคลระหว่างมุมมอง และยังระบุได้ในเบื้องต้นว่ามีบุคคลใดบ้างเข้า-ออกในบริเวณที่วิเคราะห์ ซึ่งในระบบการติดตามและจดจำตัวบุคคลที่ได้นำเสนอไปนี้จะใช้ข้อมูลตำแหน่ง และสี ซึ่งสีเป็นข้อมูลเบื้องต้นที่เด่นชัดที่สามารถ

นำมาใช้ในการแยกแยะแต่ละบุคคลทั้งในกล้องเดียวกันและระหว่างกล้อง ซึ่งมีความซับซ้อนในการประมวลผลที่น้อย นอกจากนี้งานวิจัยนี้ได้นำเสนอการตรวจจับเหตุการณ์ต่าง ๆ ที่มีความน่าสนใจ ซึ่งอาจจะเกิดขึ้นได้ในระบบเฝ้าระวังและดูแลด้านสุขภาพ ได้แก่ การตรวจจับการล้มซึ่งเป็นการประยุกต์ต่อยอดจากการวิเคราะห์ท่าทางพื้นฐาน โดยการจับลำดับการเปลี่ยนแปลงจากท่าอื่น ๆเป็นท่านอน โดยต้องใช้สถานที่มาเป็นตัวยืนยันว่าการเกิดนอนหรือการล้ม โดยในงานวิจัยนี้ใช้การสร้างจุดยกเว้น ซึ่งอาจจะเป็นที่นอน โซฟา เบาะตั้งพื้นหรือเก้าอี้ โดยที่จุดยกเว้นนั้น จะใช้วิธีการให้บุคคลเข้าไปนอนในตำแหน่งที่จะยกเว้น แล้วทำการตรวจจับตำแหน่งที่ยกเว้นมาเก็บไว้ ซึ่งถ้าหากมีการนอนในบริเวณที่ยกเว้นระบบก็จะได้ไม่ตรวจจับว่าเกิดการล้มเกิดขึ้น การตรวจจับการโบกมือเพื่อขอความช่วยเหลือ จะประยุกต์ใช้เทคนิคการวิเคราะห์ภาพฉายไบนารี่ เพื่อตรวจจับการโบกมือในแต่ละมุมมอง, และสุดท้ายเป็นส่วนของระบบการตรวจจับการกระโดด ซึ่งอาจจะเกิดได้ในกรณีของการตกใจจากการถูกจี้ปล้น หรือวิ่งหลบบางอย่าง ซึ่งเป็นเหตุการณ์ที่ไม่ปกติสำหรับในพื้นที่ร่มในอาคาร จะใช้การวิเคราะห์เลเยอร์แบบจำลองของความเคลื่อนไหวที่เปลี่ยนแปลงไปแต่ละเวลาเพื่อตรวจจับจังหวะดีดตัวขึ้นและดิ่งลงซึ่งเป็นองค์ประกอบสำคัญของการกระโดด โดยต้องผ่านการยืนยันร่วมกันกับผลของการวิเคราะห์ท่าทางที่โดยปกติต้องเป็นท่ายืน เพื่อตัดผลบวกปลอมออกไป

บทที่ 4
ผลการทดสอบ

สำหรับในผลการทดสอบจะกล่าวถึงชุดข้อมูลที่ใช้ทดสอบการรู้จำท่าทาง 3 ชุดข้อมูล ได้แก่ PSU, NW-UCLA, i3DPost ถัดมาคือส่วนของชุดข้อมูลสำหรับการติดตามและจดจำตัวบุคคล PSU และชุดข้อมูลที่ใช้ทดสอบการตรวจจับท่าทางที่ผิดปกติ ได้แก่ ชุดข้อมูลสำหรับการตรวจจับการกระโดด PSU, ชุดข้อมูลสำหรับการตรวจจับการโบกมือขอความช่วยเหลือ PSU ในส่วนถัดมาจะอธิบายถึงวิธีการทดสอบ ผลการทดสอบ และการวิเคราะห์ในหัวข้อต่าง ๆซึ่งได้แก่ การรู้จำท่าทางโดยการฟิวชันข้อมูลในระดับสูงจากหลายมุมมอง, การทดสอบการรู้จำท่าทางโดยการฟิวชันข้อมูลในระดับสูงจากหลายมุมมอง, การทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง, และการทดสอบการตรวจจับท่าทางที่ผิดปกติ

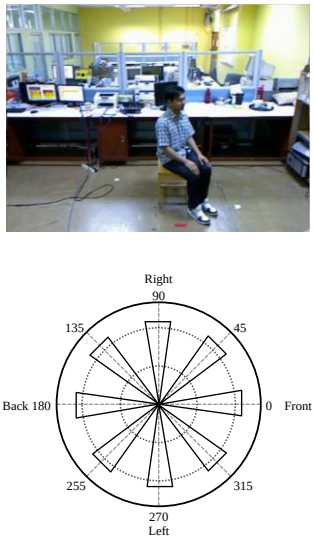
4.1 ชุดข้อมูลที่ใช้ในการทดสอบ

4.1.1 ชุดข้อมูลสำหรับการรู้จำท่าทาง PSU

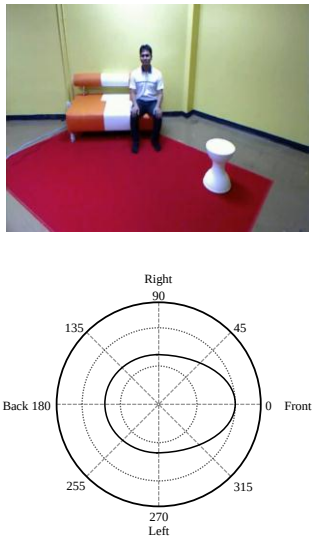
ชุดข้อมูลสำหรับการรู้จำท่าทาง PSU [96] ประกอบไปด้วยวีดีโอภาพเคลื่อนไหวของท่าทางพื้นฐานของมนุษย์ 4 ท่าทาง ซึ่งถูกบันทึกจาก 2 มุมมองที่แตกต่างกันโดยใช้กล้องสีและความลึก (RGB-D) โดยวีดีโอที่อยู่ในชุดทดสอบเดียวกันจะถูกบันทึกพร้อม ๆ กัน และมีการเชื่อมต่อประสานทางเวลาให้ตรงกัน (Frame Synchronized) ท่าทางพื้นฐานของมนุษย์ที่ได้บันทึกประกอบไปด้วย ท่ายืน/เดิน ท่านั่ง ท่าก้ม และท่านอน โดยมี 2 ฉากที่แตกต่างกัน คือ ฉากในห้องทำงานและอีกฉากหนึ่งในห้องนั่งเล่น ซึ่งได้แสดงถึงตัวอย่างในแต่ละฉากและการกระจายตัวของมุมมองตามฉากนั้น ๆ ในภาพประกอบที่ 4-1 โดยฉากในห้องทำงานจะถ่ายเฉพาะ 5 มุมมองหลัก ๆ ได้แก่ หน้า (0°), เฉียงทางหน้า (45°, 315°), ข้าง (90°,270°), เฉียงทางหลัง (135°,255°), และหลัง (180°) ส่วนฉากในห้องนั่งเล่นนั้น จะมีมุมมองที่ต่อเนื่องกันทุก ๆ มุมมอง เนื่องจากบุคคลที่อยู่ในชุดทดสอบจะเคลื่อนไหวโดยอิสระ โดยจะเน้นหนักไปทางด้านหน้า

โดยกล้องความลึกทั้งสองมุมมองจะถูก ติดตั้งไว้เหนือหัวทำมุม 60 องศา กับ แนวดิ่ง ณ ยอดสุดของเสา 2 เมตร โดยที่ข้อมูลภาพสีและความลึกจากหลายมุมมองจะถูกบันทึกจากกล้องที่อยู่นิ่ง จากหลาย ๆมุมมองที่สังเกตการณ์ในพื้นที่ที่สนใจ ซึ่งมองไปยังบริเวณเดียวกัน โดยระยะการทำงานของระบบประมาณ 3 ถึง 5.5 เมตร จากกล้องเพื่อที่จะได้ภาพเต็มตัวของมนุษย์ ดังได้แสดงไว้แล้วในภาพประกอบที่ 4-2 โดยที่ภาพสีที่ได้บันทึกมาในชุดภาพเคลื่อนไปไหว จะมีความละเอียด 640 x 480 ในขณะที่ภาพความลึกมีระดับความละเอียด 8 และ 24 บิต ซึ่งมีความ

ละเอียดที่เท่ากันกับภาพสี ในแต่ละลำดับเหตุการณ์จะมีผู้แสดง 3-5 คน และมีไม่น้อยกว่า 40 เฟรมที่เป็นพื้นหลังในตอนเริ่มต้นของลำดับเหตุการณ์ เพื่อที่จะสามารถใช้การตรวจจับความเคลื่อนไหว เพื่อแยกพื้นหลังออกจากตัวบุคคล โดยเฟรมเรทในแต่ละวีดีโอ จะมีประมาณ 8-12 เฟรมต่อวินาที

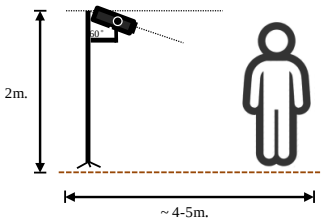


(ก) ภาพตัวอย่างและการกระจายตัวของมุมมองในฉากห้องทำงาน



(ข) ภาพตัวอย่างและการกระจายตัวของฉากในห้องนั่งเล่น

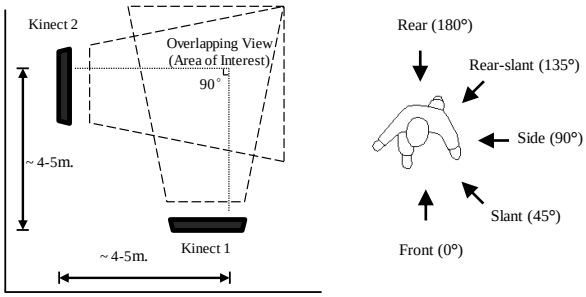
ภาพประกอบที่ 4-1 ตัวอย่างของฉากทั้งสองของชุดข้อมูล PSU



ภาพประกอบที่ 4-2 การติดตั้งกล้องเพื่อเก็บชุดข้อมูล PSU

(ก) ฉากในห้องทำงาน

จากภาพประกอบที่ 4-3 แสดงถึงแผนผังในฉากห้องทำงาน โดยที่กล้อง 2 ตัวที่ทำมุมฉากกัน โดยมุมที่กล้องชี้ไปยังบุคคลมี 5 มุมมอง ได้แก่ หน้า (0°), เฉียงทางหน้า (45, 315°), ข้าง (90,270°), เฉียงทางหลัง (135,255°), และหลัง (180°) โดยมีจำนวนของเฟรมที่กระทำท่าทาง 8,700 เฟรม ที่ไม่รวมเฟรมที่เป็นภาพพื้นหลัง

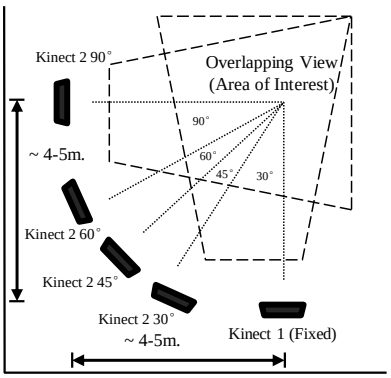


(ก) (ข)

ภาพประกอบที่ 4-3 แผนผังการติดตั้งกล้องในห้องทำงาน
(ก) แผนผังการติดตั้งจากมุมสูง (ข) มุมมองที่กล้องชี้ไปยังตัวบุคคล

(ข) ฉากในห้องนั่งเล่น

ในฉากห้องนั่งเล่นจะมี 1 กล้องที่ถูกกำหนดไว้ให้อยู่กับที่ (Kinect 1) และกล้องอีกตัวหนึ่งที่เปลี่ยนการทำมุมกับกล้องที่อยู่กับที่ (Kinect 2) โดยมีการทำมุมกัน 4 แบบ ได้แก่ ทำมุมกัน : 30°, 45°, 60°และ 90° โดยผู้แสดงท่าทางจะสามารถเคลื่อนไหวได้โดยอิสระในมุมมองต่าง ๆ ภายในพื้นที่ โดยเฟรมที่กระทำท่าทางทั้งหมด 10,720 เฟรม (ไม่รวมถึงเฟรมที่เป็นพื้นหลัง) ซึ่งจะถูกใช้ในการประเมินและทดสอบเพื่อรู้จำท่าทาง ซึ่งแสดงแผนผังของฉากในห้องนั่งเล่นตามภาพประกอบที่ 4-4



ภาพประกอบที่ 4-4 แผนผังการติดตั้งกล้องในห้องนั่งเล่น

4.1.2 ชุดข้อมูลสำหรับการรู้จำท่าทาง NW-UCLA

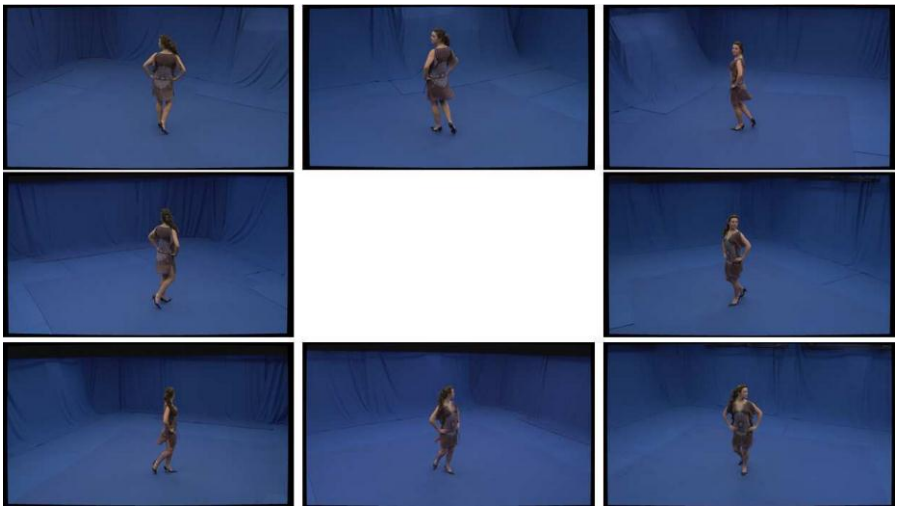
ชุดข้อมูลสำหรับการรู้จำท่าทาง NW-UCLA ที่เป็นข้อมูล 3D จากหลายมุมมอง (Northwestern-UCLA Multiview Action 3D Dataset) [69] โดยเป็นชุดข้อมูลที่ให้ข้อมูลภาพ RGB, ภาพความลึก 8 บิต ที่ทำสัญลักษณ์ส่วนที่เป็นคนด้วยสี และข้อมูลตำแหน่งของ Skeletons โดยใช้กล้อง Kinect v.1 จำนวนสามตัวจากมุมมองที่แตกต่างกันโดยที่สังเกตการณ์ไปยังพื้นที่เดียวกัน ประกอบด้วยท่าทาง 10 ท่าทาง ได้แก่ ก้มหยิบของมือเดียว, ก้มหยิบของสองมือ, ทิ้งของลงในถังขยะ, เดินไปรอบ ๆ, นั่งลง, ยืนขึ้น, สวมเสื้อ, ถอดเสื้อ, ขว้างของลงถังขยะ, และยกสิ่งของ โดยในแต่ละท่าทางจะทำโดยผู้แสดง 10 คน ดังแสดงตัวอย่างในภาพประกอบที่ 4-5

ภาพประกอบที่ 4-5 ตัวอย่างภาพชุดข้อมูลสำหรับการรู้จำท่าทาง NW-UCLA

โดยในงานวิจัยนี้จะดึงข้อมูลภาพความลึกที่เป็นท่าทางพื้นฐานของมนุษย์ ได้แก่ ท่าเดิน/ยืน ที่ถูกดึงออกมาจากท่ายืนขึ้นและท่าเดินรอบ ๆ, ท่านั่ง ที่ถูกดึงออกมาจากท่านั่งลงและท่ายืนขึ้น, ท่าก้ม ที่ถูกดึงออกมาจากท่าก้มเก็บของมือเดียวและสองมือ เพื่อให้เข้ากับท่าทางของวิทยานิพนธ์ฉบับนี้

4.1.3 ชุดข้อมูลสำหรับการรู้จำท่าทาง i3DPost

ชุดข้อมูลสำหรับการรู้จำท่าทาง i3DPost [97] เป็นชุดข้อมูลภาพสีที่แสดงท่าทางของมนุษย์ที่ได้จากการตั้งกล้อง 8 ตัวในมุมมองที่แตกต่างกันโดยมีมุม 45° ระหว่างกล้อง (ตามภาพประกอบที่ 4-6 และ 4-7) และมีการเชื่อมต่อประสานทางเวลาให้ตรงกัน (Frame Synchronized) โดย 1 จากทั้งหมด 8 คนจะทำท่าทางต่างๆจำนวน 12 ครั้ง โดยที่นักแสดงจะมีขนาดตัวที่แตกต่างกัน เสื้อผ้าที่แตกต่างกัน รวมไปถึงเพศและเชื้อชาติที่แตกต่างกันอีกด้วย ชุดข้อมูลจะถูกแบ่งเป็น 2 ชุด โดยที่ตำแหน่งแตกต่างกัน 2 ตำแหน่ง โดยแต่ละชุดจะให้ข้อมูลของภาพพื้นหลัง ซึ่งจะสามารถนำมาใช้ในการตรวจจับการเคลื่อนไหวได้ สำหรับชุดข้อมูลนี้ประกอบด้วยท่าทางดังนี้ เดิน, วิ่ง, กระโดดไปด้านหน้า, ก้ม, โบกมือ, กระโดดอยู่กับที่, นั่งแล้วยืนขึ้น, วิ่งแล้วหกล้ม, เดินแล้วมานั่ง, วิ่งแล้วกระโดดแล้วเดิน, จับมือทักทาย และ ดึงมือคืนอื่น



ภาพประกอบที่ 4-6 ตัวอย่างมุมมองของชุดข้อมูลสำหรับการรู้จำท่าทาง i3DPost [97]

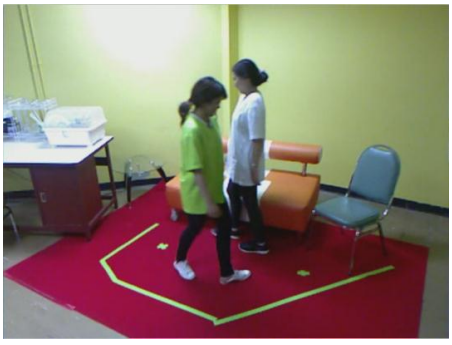
ภาพประกอบที่ 4-7 ตัวอย่างท่าทางพื้นฐานของชุดข้อมูลสำหรับการรู้จำท่าทาง i3DPost

4.1.4 ชุดข้อมูลสำหรับการติดตามและจดจำตัวบุคคล PSU

ข้อมูลที่ใช้ในการทดสอบจะเป็นฉากของห้องนั่งเล่น ให้แสงสว่างตามปกติ ซึ่งจะเก็บทั้งข้อมูลภาพสีและความลึก ที่ความละเอียด 640x480 พิกเซล โดยใช้กล้อง Kinect v.1 จาก 2 มุมมองที่แตกต่างกัน โดยนักแสดงจะใส่เสื้อพื้นที่มีสีที่แตกต่างกัน ซึ่งมีระยะที่ใช้ในการเก็บข้อมูลจะอยู่ในระยะ 3 ถึง 5.5 เมตร และจำกัดความสูงของข้อมูลจากนักแสดงที่ 180 ซม. เพื่อให้ความลึกไม่มีการสูญเสียและได้ข้อมูลที่เป็นส่วนของร่างกายได้ครบถ้วนตั้งแต่ขาถึงศีรษะ โดยชุดข้อมูลสำหรับการติดตามและจดจำตัวบุคคล PSU ที่ประกอบไปด้วยวีดีโอที่เป็นชุดข้อมูลที่ใช้ในการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละคน (Single Tracking & Re-identification) ที่มี 3 Datasets และชุดข้อมูลการติดตามและจดจำโดยเข้าไปในระบบครั้งละสองคน (Multiple Tracking & Re-identification) ที่มี 2 Datasets

โดยชุดข้อมูลการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละคน (Single Tracking & Re-identification Dataset) จะมีบุคคลที่แสดง 5 บุคคล ในแต่ละ Dataset โดยแต่ละบุคคลมีลักษณะสีเสื้อที่แตกต่างกันชัดเจน โดยทดสอบทั้งหมด 4 ท่าทาง คือ ยืนและเดิน (Standing & Walking), นั่ง (Sitting), ก้ม (Bending) และนอน (Laying) โดยจะมีข้อมูลที่เป็น Video ชุดแรกป้อนเข้าไปให้ระบบติดตามและจดจำ (Identifying Set) โดยเข้าไปครั้งละหนึ่งคน และชุดที่สองเป็นชุดการทดสอบหลังจากที่ป้อนข้อมูลชุดแรกเข้าไปเพื่อให้ระบบติดตามและจดจำ (Testing Set)

ส่วนชุดข้อมูลการติดตามและจดจำโดยเข้าไปในระบบครั้งละสองคน (Multiple Tracking & Re-identification Dataset) จะมีข้อมูลที่เป็น Video ชุดแรกป้อนเข้าไปให้ระบบติดตามและจดจำ (Identifying Set) โดยเข้าไปครั้งละหนึ่งคนเช่นเดียวกัน หลังจากนั้นจะเป็นชุดการทดสอบโดยจะเป็น Video ทดสอบโดยเข้าไปที่ละหนึ่งคนจนครบสองคน โดยจะทดสอบพิเศษเฉพาะตามท่าทางต่าง ๆในหลาย ๆมุมมองเช่นเดียวกับการทดสอบประเภทแรก โดยแต่ละชุดของข้อมูลจะเริ่มจากการไปยืนตามตำแหน่งที่จะไม่มีการบดบังเพื่อให้ระบบสามารถติดตามและจดจำได้ก่อนจึงเริ่มเปลี่ยนท่าทางในแต่ละการทดสอบ โดยระบบจะนับความถูกต้องเฉพาะท่าทาง ๆนั้น ๆที่ทำการทดสอบ โดยจะมีการเดินเพื่อสลับตำแหน่งที่ทดสอบทั้งหมด 3 ครั้ง ตามภาพประกอบที่ 4-8

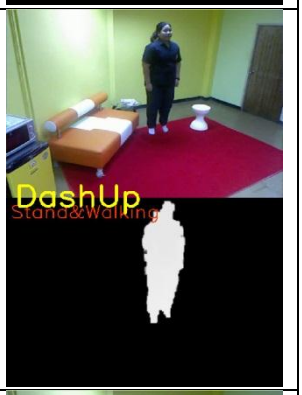
ภาพประกอบที่ 4-8 ตัวอย่างภาพในชุดข้อมูลการติดตามและจดจำโดยเข้าไปในระบบครั้งละหนึ่งและสองคน

4.1.5 ชุดข้อมูลสำหรับการตรวจจับการกระโดด PSU

สำหรับชุดข้อมูลสำหรับการตรวจจับการกระโดด จะเก็บทั้งข้อมูลภาพสีและความลึก ที่ความละเอียด 640x480 พิกเซล 2 มุมมองที่แตกต่างกัน จากกล้อง Kinect v.1 โดยจะเป็นฉากของห้องนั่งเล่น ซึ่งมีระยะอยู่ที่ 3 ถึง 5.5 เมตร จากกล้อง และจำกัดความสูงของผู้กระโดดที่ 180 ซม. เพื่อให้ความลึกไม่มีการสูญเสียและได้ข้อมูลที่เป็นส่วนของร่างกายได้ครบถ้วนตั้งแต่ขาถึงศีรษะ

โดยชุดข้อมูลการกระโดดจะแบ่งเป็น 2 ประเภท คือ กระโดดลงที่เดิม และกระโดดลงอีกที่ โดยในแต่ละประเภทจะมีทั้งการกระโดดสูง และการกระโดดต่ำๆ โดยจะใช้ผู้ทดสอบ 6 คน เป็นผู้ชาย 3 คน และผู้หญิง 3 คน ที่มีรูปร่างและความสูงต่างกัน ดังแสดงลักษณะของผู้ที่เป็นทดสอบดังตารางที่ 4-1 โดยในการกระโดดแต่ละรูปแบบการทดสอบจะให้กระโดดกระโดดต่ำๆ 3 ครั้ง และกระโดดสูง 3 ครั้ง โดยมีรูปแบบการกระโดดดังนี้ กระโดดลงที่เดิมมุมมองที่ 1, กระโดดลงที่เดิมมุมมองที่ 2, กระโดดลงอีกที่มุมมองที่ 1, และกระโดดลงอีกที่มุมมองที่ 2 โดยแต่ละรูปแบบจะกระโดด 6 ครั้ง (กระโดดต่ำๆ 3 ครั้ง และกระโดดสูง 3 ครั้ง) ต่อ 1 คน รวมจำนวนกระโดดต่อ 1 คนคือ 24 ครั้ง จะได้จำนวนครั้งของการกระโดดทั้งหมด 144 ครั้ง และมีเหตุการณ์ที่ไม่ใช่การกระโดดเพื่อทดสอบความจำเพาะของระบบคือการยืนก่อนกระโดด 3 ครั้ง และรวมถึงการเปลี่ยนท่าก่อนการกระโดดอีกครั้ง 3 ครั้ง รวมเป็น 6 ครั้งในแต่ละรูปแบบการทดสอบ

ตารางที่ 4-1 ลักษณะของผู้ทดสอบการตรวจจับการกระโดด

ผู้ทดสอบที่	ลักษณะ	ผู้ทดสอบที่	ลักษณะ
1		4	
2		5	
3		6	

4.1.6 ชุดข้อมูลสำหรับการตรวจจับการโบกมือขอความช่วยเหลือ PSU

ข้อมูลที่ใช้ในการทดสอบจะเป็นฉากของห้องนั่งเล่น ให้แสงสว่างตามปกติ ซึ่งจะเก็บทั้งข้อมูลภาพสีและความลึก ที่ความละเอียด 640x480 พิกเซล โดยใช้กล้อง Kinect v.1 จาก 2 มุมมองที่แตกต่างกัน ซึ่งมีระยะที่ใช้ในการเก็บข้อมูลจะอยู่ในระยะ 3 ถึง 5.5 เมตร และจำกัดความสูงของข้อมูลจากนักแสดงที่ 180 ซม. เพื่อให้ความลึกไม่มีการสูญเสียและได้ข้อมูลที่เป็นส่วนของร่างกายได้ครบถ้วนตั้งแต่ขาถึงศีรษะ โดยชุดข้อมูลสำหรับการตรวจจับการโบกมือขอ

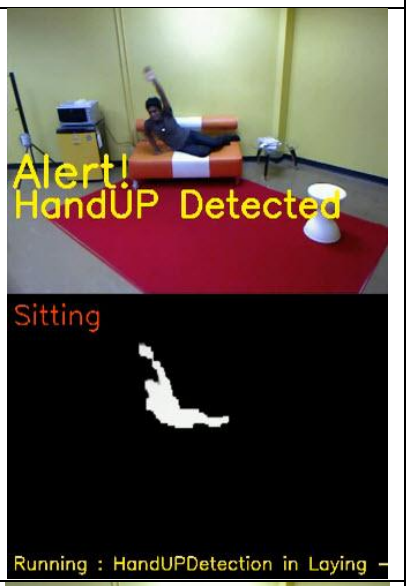
ความช่วยเหลือ โดยจะเป็นการโบกมือในท่าต่าง ๆ คือ ยืน/เดิน นั่ง ก้ม และนอน มีผู้แสดง 4 คน แต่ละคนจะโบกมือในมุมมองและตำแหน่งที่ต่างกัน ดังแสดงตัวอย่างตามตารางที่ 4-2

ตารางที่ 4-2 ตัวอย่างการโบกมือขอความช่วยเหลือตามท่าทางต่าง ๆ

ท่าทาง	ตัวอย่างที่ 1	ตัวอย่างที่ 2
ยืน		
นั่ง		
ก้ม		
นอน		

ในชุดทดลองตรวจจับการโบกมือจะแบ่งเป็นการโบกมือในท่าทางพื้นฐานต่าง ๆของมนุษย์ ได้แก่ ยืน / เดิน นั่ง นอน และก้ม และทำซ้ำในหลาย ๆมุมมอง โดยในการโบกมือ 1 เหตุการณ์ถ้ามีการตรวจจับได้ 1 ครั้งให้ถือว่าระบบตรวจจับได้ โดยได้ทำการทดสอบโดยใช้ข้อมูลจากผู้ทดสอบจำนวน 4 คน ในแต่ละคนต้องโบกมือใน 1 ท่าทาง 8 ครั้ง รวมต่อคนต้องโบกมือ 32 ครั้ง โดยในแต่ละท่าจะมีการโบกมือ 32 ครั้งเช่นกัน โดยจำนวนครั้งการทดสอบทั้งหมดมี 128 ครั้ง ซึ่งได้แสดงลักษณะของผู้ทดสอบการตรวจจับการโบกมือไว้ในตารางที่ 4-3

ตารางที่ 4-3 ลักษณะของผู้ทดสอบการตรวจจับการโบกมือ

ผู้ทดสอบที่	ลักษณะ	ผู้ทดสอบที่	ลักษณะ
1		3	
2		4	

4.2 การทดสอบการรู้จำท่าทางโดยการฟิวชันข้อมูลในระดับสูงจากหลายมุมมอง

4.2.1 การทดสอบโดยชุดข้อมูล PSU

สำหรับการทดสอบจะใช้ข้อมูลในการทดสอบเป็นฉากของห้องทำงาน ในชุดข้อมูล PSU โดยใช้คอมพิวเตอร์ที่มีหน่วยประมวลผล CPU Intel Core i7 4700MQ ที่ความถี่ 2.40GHz และหน่วยความจำหลัก DDR3 8GB โดยทดสอบจำนวนประมาณ 8700 เฟรมโดยมีสองมุมมองที่ตั้งฉากกันและสังเกตการณ์ไปยังบริเวณที่ซ้อนทับกัน โดยการทดสอบใช้เวลาในการประมวลผล 110 ms สำหรับการรู้จำท่าทางมุมมองเดี่ยว, 240 ms สำหรับการฟิวชันระดับสูงหลายมุมมองโดยไม่ใช้การประมวลผลแบบขนาน โดยที่ใช้เวลา 180 ms สำหรับการประมวลผลแบบขนาน โดยการทดสอบจะแบ่งเป็นสองส่วนตามแบบจำลองการฟิวชัน ได้แก่ แบบจำลองฟิวชันเบื้องต้นและซับซ้อน โดยผลลัพธ์ของการรู้จำท่าทางในมุมมองเดี่ยวได้แสดงตามตารางที่ 4-4

ตารางที่ 4-4 ความแม่นยำของการรู้จำในท่าทางและมุมมองต่าง ๆในมุมมองเดี่ยว

ท่าทาง	ความแม่นยำตามมุมมอง (%)					
	หน้า	เฉียงทางหน้า	ข้าง	เฉียงทางหลัง	หลัง	ค่าเฉลี่ย
ยืน / เดิน	81.21	76.59	75.93	72.86	61.83	73.68
นั่ง	85.61	44.75	53.72	29.97	93.00	61.41
ก้ม	62.57	73.61	85.80	3.95	6.32	46.45
นอน	0	78.67	98.06	89.56	0	53.25

(ก) แบบจำลองฟิวชันเบื้องต้น

จากผลการทดสอบแบบจำลองฟิวชันเบื้องต้นตามตารางที่ 4-5 ให้ผลลัพธ์ความแม่นยำมากที่สุดถึง 98.71% ในท่านอนด้านข้าง แต่ค่อนข้างทำได้ดีน้อยกว่าที่ควรในท่านั่ง ซึ่งท่านั่งได้ตอบผิดพลาดไปเป็นท่ายืนถึง 23.75% โดยแบบจำลองนี้ทำให้ความแม่นยำโดยเฉลี่ยเพิ่มมากขึ้นเป็น 11.86% เมื่อเทียบกับมุมมองเดี่ยว

ตารางที่ 4-5 ผลการทดสอบความแม่นยำของแบบจำลองฟิวชันเบื้องต้น

ท่าทาง	ความแม่นยำตามมุมมอง (%)					
	หน้า	เฉียงทางหน้า	ข้าง	เฉียงทางหลัง	หลัง	ค่าเฉลี่ย
ยืน / เดิน	83.54	79.90	87.97	84.17	67.98	80.71
นั่ง	68.87	26.10	68.58	17.67	88.48	53.94
ก้ม	82.34	60.05	84.66	59.32	39.12	65.10
นอน	77.66	96.89	98.71	94.95	44.22	82.49

(ข) แบบจำลองฟิวชันซับซ้อน

ในการทดสอบแบบจำลองฟิวชันซับซ้อนได้ใช้ค่า ซึ่งแสดงผลลัพธ์ตามมุมมองและท่าทางต่าง ๆ ในตารางที่ 4-6 ซึ่งผู้วิจัยพบว่าแบบจำลองฟิวชันซับซ้อนสามารถเพิ่มความแม่นยำอย่างมีนัยสำคัญให้กับท่าก้มในด้านหน้า และท่านอนในด้านข้างถึง 99.40% และ 98.39% ตามลำดับ แต่ก็ทำให้ความแม่นยำของท่ายืน/เดิน และนั่ง ลดลงเล็กน้อย โดยให้ผลลัพธ์ที่ต่ำในท่ายืน/เดิน ในขณะที่ท่าทางอื่นมีแนวโน้มที่ดีขึ้น เนื่องจากว่าแบบจำลองนี้จะทำให้ท่ายืน/เดิน ให้คำตอบผิดไปเป็นท่าอื่นเป็นส่วนมาก โดยแบบจำลองนี้ทำให้ความแม่นยำโดยเฉลี่ยเพิ่มมากขึ้นเป็น 16.66% เมื่อเทียบกับมุมมองเดี่ยว ซึ่งตัวอย่างการรู้จำท่าทางโดยการฟิวชันระดับสูงได้แสดงในตารางที่ 4-7

ตารางที่ 4-6 ผลการทดสอบความแม่นยำของแบบจำลองฟิวชันซับซ้อน

ท่าทาง	ความแม่นยำตามมุมมอง (%)					
	หน้า	เฉียงทางหน้า	ข้าง	เฉียงทางหลัง	หลัง	ค่าเฉลี่ย
ยืน / เดิน	51.44	59.80	67.74	69.80	56.22	61.00
นั่ง	67.22	28.30	66.33	54.09	75.00	58.19
ก้ม	99.40	90.96	98.30	90.69	52.39	86.34
นอน	98.17	99.12	98.39	97.98	85.81	95.89

ตารางที่ 4-7 ตัวอย่างการทดสอบการรู้จำท่าทางโดยการฟิวชันข้อมูลในระดับสูงจากหลายมุมมอง

ตัวอย่างที่ / ท่าทาง	ตัวอย่างการทดสอบการรู้จำท่าทางโดยการฟิวชันข้อมูลในระดับสูง แบบจำลองฟิวชันซับซ้อน
1	

ตัวอย่างที่ / ท่าทาง	ตัวอย่างการทดสอบการรู้จำท่าทางโดยการฟิวชันข้อมูลในระดับสูงแบบจำลองฟิวชันซับซ้อน
2	
3	
4	

ตัวอย่างที่ / ท่าทาง	ตัวอย่างการทดสอบการรู้จำท่าทางโดยการฟิวชันข้อมูลในระดับสูง แบบจำลองฟิวชันซับซ้อน
5	

4.2.2 วิเคราะห์ผลการทดสอบการรู้จำท่าทางโดยการฟิวชันข้อมูลในระดับสูง

ในงานวิจัยนี้ได้นำเสนอการรู้จำท่าทางโดยการฟิวชันข้อมูลในระดับสูงจากหลายมุมมองที่มีเป้าหมายเพื่อการเพิ่มความแม่นยำในการรู้จำท่าทาง ซึ่งจากผลการทดสอบการรู้จำท่าทางโดยการฟิวชันปรากฏว่ามีความแม่นยำเพิ่มมากขึ้น 11.86% และ 16.66% ของฟิวชันแบบจำลองพื้นฐานและซับซ้อนตามลำดับ โดยสามารถเพิ่มค่าความแม่นยำในท่านอนจากมุมมองด้านหน้ามากที่สุดถึง 98.17% จากมุมมองเดี่ยว ซึ่งทั้งสองฟิวชันแบบจำลองจะให้ผลลัพธ์ที่แตกต่างกัน โดยแบบจำลองซับซ้อนจะให้ความแม่นยำโดยเฉลี่ยในภาพรวมเล็กน้อย โดยสามารถเพิ่มความแม่นยำให้กับท่าก้มท่านอนอย่างมีนัยสำคัญ แต่ก็ทำให้ความแม่นยำของท่ายืน/เดิน และนั่ง ลดลงเล็กน้อย

4.3 การทดสอบการรู้จำท่าทางระดับการฟิวชันฟีเจอร์ในระดับล่างจากหลายมุมมอง

สำหรับการทดสอบประสิทธิภาพของการรู้จำท่าทางจากแบบจำลองเลเยอร์แบบฟิวชันข้อมูลภาพสีและความลึกจากหลายมุมมองจะถูกทดสอบในชุดข้อมูล 3 ชุดข้อมูล ได้แก่ PSU (Prince of Songkla University) NW-UCLA (Northwestern-University of California at Los Angeles) และ i3DPost ซึ่งผู้วิจัยได้ใช้ชุดข้อมูล PSU เพื่อประมาณค่าพารามิเตอร์ต่าง ๆของแบบจำลองของนี้ อย่างเช่น จำนวนของเลเยอร์, พารามิเตอร์ α ที่ปรับได้, และทดสอบประสิทธิภาพในเงื่อนไขต่าง ๆ เช่น มุมมองจากกล้องเดี่ยวและหลายมุมมองจากหลายกล้อง องศาที่กระทำกันระหว่างกล้อง และวิธีการที่ใช้ในการรู้จำถัดจากนั้น วิธีการของผู้วิจัยจะถูกทดสอบโดยใช้ชุดข้อมูล NW-UCLA และ i3DPost ซึ่งถูกตั้งที่มุมมองที่ต่างกัน รวมไปถึงองศาระหว่างกล้องเพื่อที่จะใช้ในการประเมินความทนทานของแบบจำลองของผู้วิจัย โดยการทดสอบการรู้จาท่าทางระดับการฟิวชันฟีเจอร์ในระดับล่างทั้งหมดจะใช้คอมพิวเตอร์ที่ใช้หน่วยประมวลผล Intel Core i5 4590 at 3.30GHz และ DDR3 8GB ส่วนในวิธีการจำแนกประเภทข้อมูลของโครงข่ายประสาทเทียมจะใช้ 1 เลเยอร์ซ่อน -20 โหนด ด้วย Sigmoid Function ที่รู้จำโดยการ Feed-forward และสอนโมเดลโดย Back-propagation ส่วนซัพพอร์ตเวกเตอร์แมชชีน จะใช้ Radial Basis Function Kernel ร่วมกันกับ C-SVC

4.3.1 การทดสอบโดยชุดข้อมูล PSU

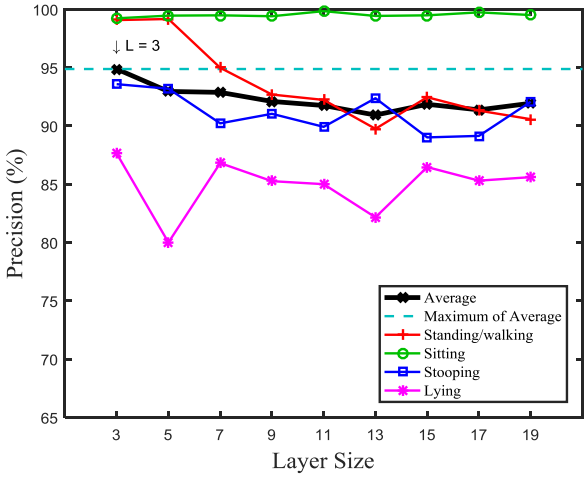
สำหรับทดสอบการรู้จำท่าทางระดับการฟิวชันฟีเจอร์ในระดับล่างจากหลายมุมมองด้วยชุดข้อมูล PSU จะใช้ฉากในห้องทำงานเพื่อใช้สำหรับสอนข้อมูลตัวเรียนรู้เพื่อรู้จำ และอีกฉากหนึ่งในห้องนั่งเล่นสำหรับใช้ในการทดสอบประสิทธิภาพ

(ก) การประเมินของค่าจำนวนของเลเยอร์

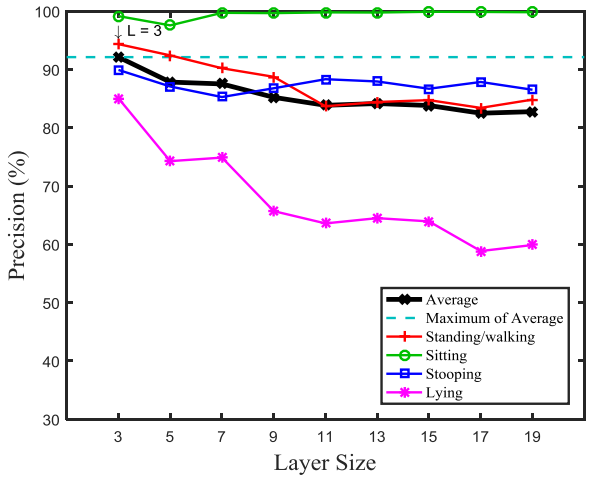
ผู้วิจัยได้ทำการประเมินค่าของจำนวนเลเยอร์ (L) ที่เหมาะสมโดยการทดสอบแบบจำลองของผู้วิจัยโดยที่มีค่าจำนวนของเลเยอร์ที่แตกต่างกัน โดยใช้ชุดข้อมูล PSU โดยจำนวนของเลเยอร์ที่ได้ทดสอบมีดังนี้ 3,5,7,9,11,13,15,17 และ 19 โดยที่กำหนดค่า α เท่ากับ 0.7 โดยใช้วิธีการแยกแยะ เพื่อเรียนรู้และทดสอบการรู้จำ 2 ตัว คือ โครงข่ายประสาทเทียม (ANN) และ Support Vector Machine (SVM) โดยผลลัพธ์ของทั้งสอง โดยมีจำนวนเลเยอร์ที่แตกต่างกัน ดังแสดงไว้ในภาพประกอบที่ 4-9 และ 4-10 ตามลำดับ

โดยที่ภาพประกอบที่ 4-9 แสดงให้เห็นว่าจำนวนของเลเยอร์ที่มีขนาดเท่ากับ 3 ประสบความสำเร็จสูงสุด โดยมีค่าเฉลี่ยความแม่นยำมากที่สุดที่ 94.88% โดยใช้ ANN และ 92.11% โดยใช้ SVM โดยแสดงไว้ในภาพประกอบที่ 4-10 โดยที่ ANN ทำได้ดีกว่าเล็กน้อย ซึ่งต่อจากนี้

การทดสอบของผู้วิจัยในชุดข้อมูลของ PSU จะใช้จำนวนเลเยอร์ที่มีขนาด 3 เลเยอร์ และใช้ ANN ในการประเมินและทดสอบอื่น ๆต่อไปเท่านั้น



ภาพประกอบที่ 4-9 ความแม่นยำของการทดสอบรู้จำท่าทางในจำนวนเลเยอร์ที่แตกต่างกันตามท่าทาง โดยใช้โครงข่ายประสาทเทียม (ANN) เป็นตัวเรียนรู้และทดสอบการรู้จำ

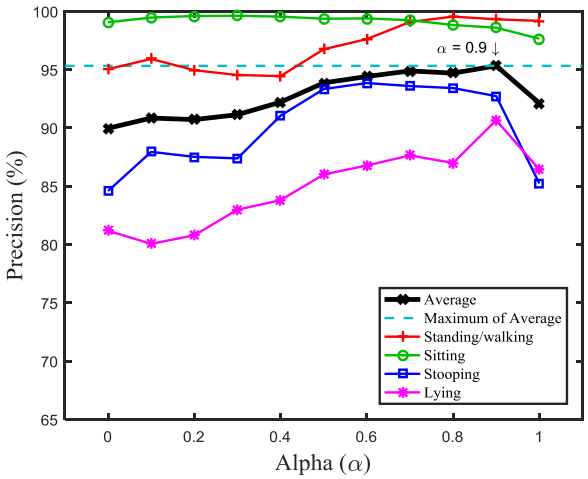


ภาพประกอบที่ 4-10 ความแม่นยำของการทดสอบรู้จำท่าทางในจำนวนเลเยอร์ที่แตกต่างกันตามท่าทาง โดยใช้ Support Vector Machine (SVM) เป็นตัวเรียนรู้และทดสอบการรู้จำ

(ข) การประเมินของค่าอัตราการเรียนรู้ α

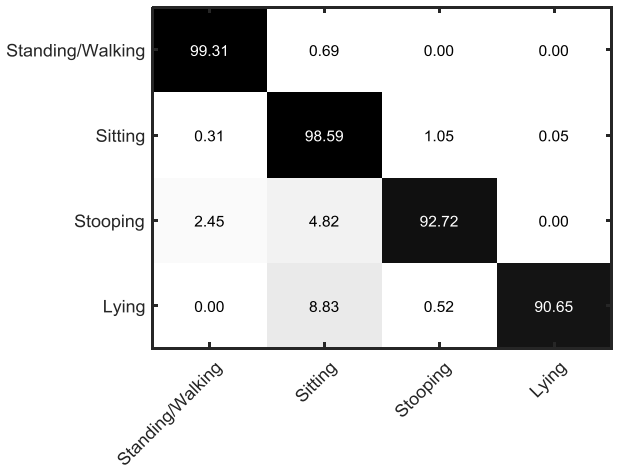
ค่าอัตราการเรียนรู้ α ซึ่งเป็นพารามิเตอร์ที่สามารถปรับค่าได้ โดยจะทำให้สามารถถ่วงค่าระหว่าง $Q[k]$ และ $\rho[k]$ เพื่อการปรับตั้งค่าได้ เมื่อค่าความหนาแน่นที่ถ่วงน้ำหนักโดยค่าย้อนกลับของความลึก $Q[k]$ มาก ค่าความหนาแน่น ($\rho[k]$) จะน้อยลงตามลำดับตามสัดส่วนที่หัก

ลบกันใน 100 ส่วน จะใช้เป็นตัวกำหนดค่าของ***ค่าถ่วงน้ำหนักความหนาแน่นโดยความลึกของเลเยอร์*** (Z[k]) ซึ่งการปรับค่าจะเป็นปรับตัวแปรที่จะเพิ่มลดความสามารถในการรู้จำของท่าทางบางส่วนโดยใช้ ค่าย้อนกลับความลึกที่อยู่ใน Q[k] เพื่อที่จะวัดปริมาตรที่ซ่อนอยู่ภายในโครงสร้างของตัวบุคคล โดยในการทดสอบประเมินค่าจะใช้ค่าตั้งแต่ 0 ถึง 1 โดยมีระยะห่างที่ 0.1 รวมทั้งหมด 11 ค่า ดังแสดงไว้ในภาพประกอบที่ 4-11



ภาพประกอบที่ 4-11 ความแม่นยำของการทดสอบรู้จำท่าทางในค่าอัตราการเรียนรู้ α ที่แตกต่างกันในแต่ละท่าทางและค่าเฉลี่ย โดยใช้ L=3 และ ANN

จากภาพประกอบที่ 4-11 ซึ่งแสดงค่าความแม่นยำของการทดสอบรู้จำท่าทางในค่าอัตราการเรียนรู้ α ที่แตกต่างกันในแต่ละท่าทางและค่าเฉลี่ย โดยใช้ L=3 และ ANN ซึ่งโดยทั่วไปยกเว้นท่านั่ง สังเกตได้ว่าการเพิ่มสัดส่วนของค่าความหนาแน่นที่ถ่วงน้ำหนักโดยค่าย้อนกลับของความลึก Q[k] มากขึ้นความถูกต้องก็ยิ่งมากขึ้นตามไปด้วย โดยมีค่าความแม่นยำสูงสุดเฉลี่ยอยู่ที่ 95.32% ณ α เท่ากับ 0.9 ซึ่งหมายความว่าค่าความหนาแน่นที่ถ่วงน้ำหนักโดยค่าย้อนกลับของความลึก Q[k] เท่ากับ 90% และค่าความหนาแน่น (ρ[k]) เท่ากับ 10% ซึ่งเป็นสัดส่วนที่เหมาะสม เมื่อค่า α มากกว่า 9 ค่าความแม่นยำจะลดลง แต่สำหรับท่านั่งจะมีแนวโน้มที่แตกต่างจากตัวอื่น คือ ลดลงบ้างเล็กน้อย แต่จะเข้าใกล้ค่าความแม่นยำที่มากเสมอ

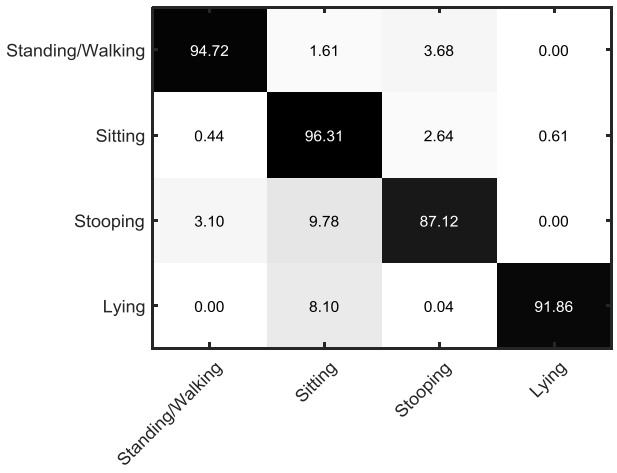


ภาพประกอบที่ 4-12 Confusion Matrix ของการรู้จำในหลายมุมมองในชุดข้อมูล PSU โดยใช้ L=3, α =0.9, และ ANN

จากภาพประกอบที่ 4-12 แสดงให้เห็นถึง Confusion Matrix ของการรู้จำในหลายมุมมองในชุดข้อมูล PSU โดยใช้ L=3, α =0.9, และ ANN ซึ่งจะพบว่า ท่ายืน/เดินมีความแม่นยำสูงสุดที่ 99.31% ในขณะที่ท่านอนมีความแม่นยำเพียงแค่ 90.65% ซึ่งความผิดพลาดเกิดขึ้นส่วนใหญ่มาจากคุณลักษณะของแกนในร่างกายของบุคคลเป็นแนวนอน ซึ่งขัดกับแบบจำลองฟีเจอร์ซึ่งออกแบบมารู้จำท่าทางในแนวดิ่ง โดยทั่วไปแล้วความผิดพลาดของท่ายืน/เดิน, ท่าก้ม, และนอน จะสับสนตอบผิดเป็นท่านั่ง 0.69%, 4.82%, และ 8.83% ตามลำดับ แต่กระนั้นค่าความแม่นยำของท่านั่งก็ยังคงสูงอยู่ที่ 98.59%

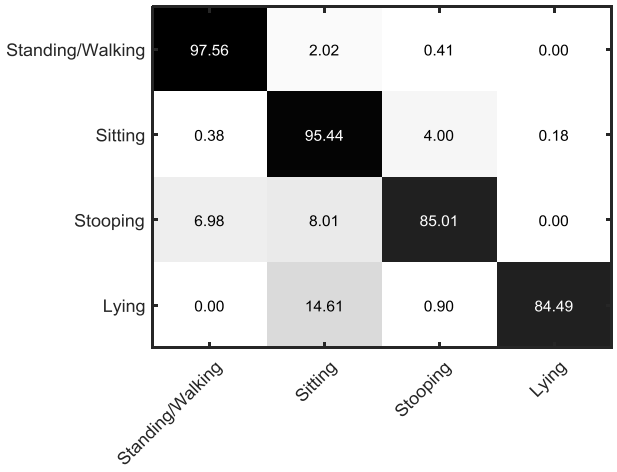
(ค) การเปรียบเทียบระหว่างการรู้จำโดยใช้มุมมองเดียวและหลายมุมมอง

เพื่อเป็นการยืนยันข้อสันนิษฐานผู้วิจัยได้ทดสอบทั้งการรู้จำในมุมมองเดียวและหลายมุมมองกับชุดข้อมูล PSU L=3, α =0.9, และ ANN ซึ่งทั้งสองใช้ชุดสอนข้อมูลจากฉากห้องทำงานเหมือนกัน โดยภาพประกอบที่ 4-13 แสดงถึงผลลัพธ์ของการรู้จำท่าทางจากมุมมองกล้องตัวที่ 1 (ซึ่งเป็นกล้องที่อยู่กับที่) ขณะที่ภาพประกอบที่ 4-14 แสดงผลลัพธ์ของมุมมองเดี่ยวกล้องที่ 2 (ซึ่งเป็นกล้องที่เปลี่ยนมุมได้)

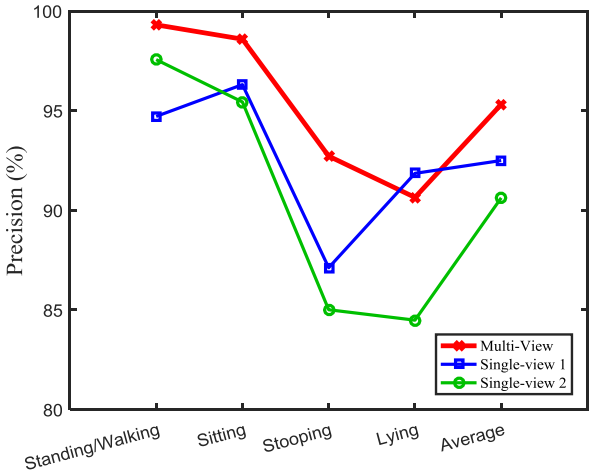


ภาพประกอบที่ 4-13 Confusion Matrix ของการรู้จำในมุมมองเดี่ยวจากกล้องที่ 1 ในชุดข้อมูล PSU โดยใช้ L=3, α =0.9, และ ANN

ผลลัพธ์แสดงให้เห็นว่ามุมมองเดี่ยวจากกล้องที่ 1 ซึ่งเป็นกล้องที่ติดกับที่มีประสิทธิภาพในการรู้จำดีกว่ามุมมองเดี่ยวจากกล้องที่ 2 เล็กน้อย (ค่าเฉลี่ยของความแม่นยำเท่ากับ 92.50% เมื่อเปรียบเทียบกับกล้องที่ 2 เท่ากับ 90.63%) มุมมองเดี่ยวจากกล้องที่ติดอยู่กับที่ให้ผลลัพธ์ที่ดีที่สุดในท่านั่ง ในขณะที่กล้องที่ 2 ที่เปลี่ยนมุมได้ให้ผลลัพธ์ดีที่สุดในท่ายืน/เดิน ซึ่งสังเกตได้ว่าการที่กล้องอยู่กับที่จะให้ผลดีกว่ากล้องที่เปลี่ยนมุมได้ในท่านอน



ภาพประกอบที่ 4-14 Confusion Matrix ของการรู้จำในมุมมองเดี่ยวจากกล้องที่ 2 ในชุดข้อมูล PSU โดยใช้ L=3, α =0.9, และ ANN

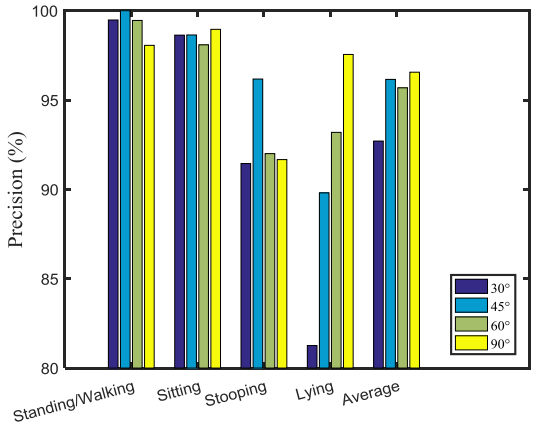


ภาพประกอบที่ 4-15 การเปรียบเทียบความแม่นยำจากหลายมุมมอง และมุมมองเดี่ยวทั้งสองกล้องโดยใช้ L=3, α =0.9, และ ANN

จากภาพประกอบที่ 4-15 แสดงให้เห็นถึงความแม่นยำของแต่ละท่าทางทั้ง 4 ท่าทางรวมไปถึงค่าเฉลี่ยทั้งหลายมุมมองและมุมมองเดี่ยวทั้งสองกล้อง โดยเฉลี่ยแล้วค่าความแม่นยำของการรู้จำจากหลายมุมมองจะให้ผลลัพธ์ที่ดีที่สุด ยกเว้นในท่านอนของมุมมองเดี่ยวจากกล้องตัวที่ 1 จะดีกว่าเล็กน้อย ซึ่งอาจจะเป็นผลมาจากการที่มุมมองเดี่ยวจากกล้องตัวที่ 1 เป็นกล้องที่ติดกับที่ จะทำให้เกิดรูปแบบของฟีเจอร์แบบเดิม ๆ ไม่เปลี่ยนแปลง ซึ่งจะทำให้การรู้จำดีกว่า

(ง) การเปรียบเทียบการทำมุมกันของกล้องที่องศาแตกต่างกัน

จากภาพประกอบที่ 4-4 ที่ได้แสดงไว้ก่อนหน้านี้ มุมมองเดี่ยวจากกล้องที่ 1 ซึ่งเป็นกล้องที่ติดกับที่ ในขณะที่มุมมองเดี่ยวจากกล้องที่ 2 เป็นกล้องที่เคลื่อนที่ได้โดยสามารถปรับมุมระหว่าง 2 กล้องได้ดังนี้ 30°, 45°, 60° และ 90° ผู้วิจัยได้ทำการทดสอบในมุมที่แตกต่างกันเหล่านี้เพื่อวัดความทนทานของแบบจำลองโดยทดสอบในท่าทางทั้ง 4 ซึ่งผลลัพธ์ได้แสดงไว้ในภาพประกอบที่ 4-16

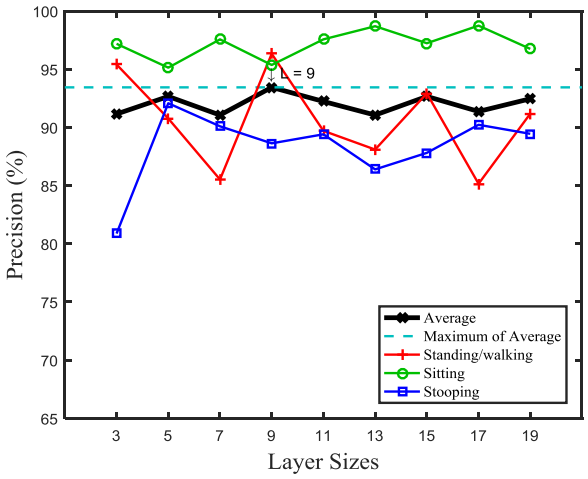


ภาพประกอบที่ 4-16 การเปรียบเทียบการทำมุมกันของกล้องที่องศาแตกต่างกัน

จากภาพประกอบที่ 4-16 ค่าความแม่นยำโดยเฉลี่ยที่ต่ำสุดเกิดขึ้น ณ มุมกล้องที่กระทำต่อกันที่ 30° ซึ่งเป็นการตั้งค่าที่เล็กที่สุดระหว่างกล้อง โดยอาจจะเป็นไปได้ว่าเกิดจากมุมที่แคบเกินไป ทำให้ทั้ง 2 กล้องได้ข้อมูลที่ค่อนข้างใกล้เคียงกัน โดยมุมอื่น ๆ ที่กล้องกระทำต่อกันมีผลลัพธ์ที่ใกล้เคียงกันโดยเฉลี่ย โดยรวมแล้วท่ายืน/เดิน และท่านั่งจะให้ผลลัพธ์ค่อนข้างใกล้เคียงกันในทุกมุมมอง ในขณะที่ท่านอนและท่าก้มจะมีผลต่อการเปลี่ยนมุมกล้องที่กระทำต่อกัน

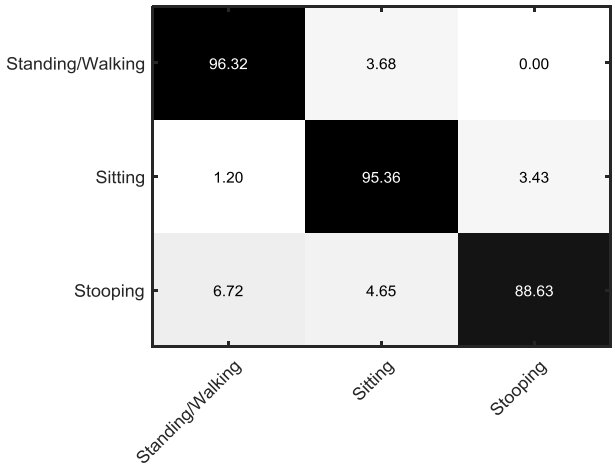
(จ) ทดสอบกับแบบจำลองที่สอนข้อมูลโดย NW-UCLA

นอกเหนือจากนี้ผู้วิจัยได้ทดสอบชุดข้อมูล PSU ในฉากห้องอยู่อาศัยโดยใช้แบบจำลองที่สอนข้อมูลโดย NW-UCLA [65] เพื่อวัดประสิทธิภาพของแบบจำลองในการทนทานต่อข้อมูลที่มีลักษณะแตกต่างกัน โดยในภาพประกอบที่ 4-17 ได้แสดงว่าการใช้ขนาดเลเยอร์ที่เท่ากับ 9 ให้ความแม่นยำที่สูงที่สุดจากค่าเฉลี่ยในทุก ๆท่าทางที่ 93.44% โดยท่านั่งให้ผลลัพธ์ที่ดีที่สุดถึง 98.74% เมื่อใช้ขนาดเลเยอร์ที่เท่ากับ 17 ในขณะเดียวกันก็ให้ผลที่ดีในเลเยอร์ที่ต่ำเช่น เลเยอร์เท่ากับ 3 มีความแม่นยำ 97.18% ในขณะที่ท่ายืน/เดิน ให้ความแม่นยำสูงสุดที่ 95.40% และต่ำสุดที่ 85.16% ส่วนท่าก้มเป็นท่าที่มีความแม่นยำน้อยที่สุดอยู่ที่ 92.08% ที่เลเยอร์ที่เท่ากับ 5



ภาพประกอบที่ 4-17 ความแม่นยำของการทดสอบรู้จำท่าทางในจำนวนเลเยอร์ที่แตกต่างกันของทุกท่าทางในชุดข้อมูล PSU ซึ่งใช้แบบจำลองที่สอนข้อมูลโดย NW-UCLA; ANN

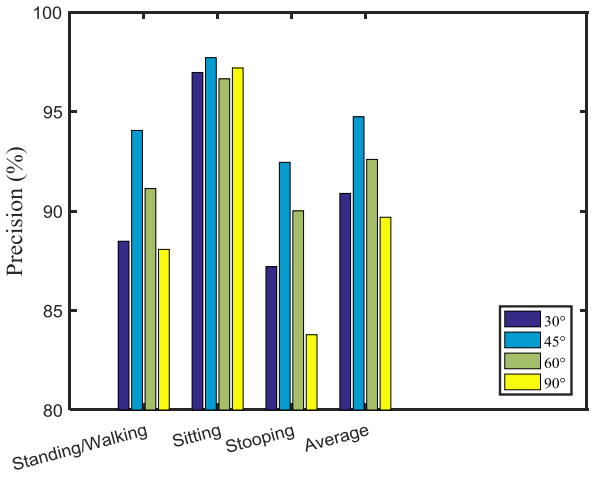
ภาพประกอบที่ 4-18 แสดงผล Confusion Matrix เมื่อจำนวนเลเยอร์เท่ากับ 9 โดยท่ายืนและท่าเดิน ให้ค่าความแม่นยำสูงสุด ที่ 96.32% ในขณะที่ท่าก้มให้ความแม่นยำน้อยสุดที่ 88.63%



ภาพประกอบที่ 4-18 Confusion Matrix ของการรู้จำสองมุมมอง ในชุดข้อมูล PSU ซึ่งใช้แบบจำลองที่สอนข้อมูลโดย NW-UCLA โดยใช้ L=9, ANN

นอกจากนี้ผู้วิจัยได้เปรียบเทียบความแม่นยำจากมุมที่กระทำต่อกันของกล้องที่แตกต่างกัน ซึ่งแสดงในภาพประกอบที่ 4-19 ผลลัพธ์ที่ดีที่สุดอยู่ที่มุม 45° ที่ 94.74% และต่ำที่สุดที่ 89.69% ที่มุม 90° ซึ่งโดยทั่วไปแล้ว ความแม่นยำของทุกท่าทางจะสูงสุดที่ 45° และลดลงเมื่อ

กว้างขึ้น อย่างไรก็ตามผลลัพธ์ที่ได้โดยใช้แบบจำลองที่สอนข้อมูลโดย PSU ให้ผลที่แตกต่างคือ เมื่อมุมมองกว้างขึ้นจะมีความแม่นยำที่มากกว่า



ภาพประกอบที่ 4-19 การเปรียบเทียบการทำมุมกันของกล้องที่องศาแตกต่างกัน ซึ่งใช้แบบจำลองที่สอนข้อมูลโดย NW-UCLA

(ฉ) การทดสอบประสิทธิภาพในด้านความเร็ว

การทดสอบประสิทธิภาพในด้านความเร็ว (จะไม่รวมระยะเวลาในการติดต่อประสานงานและแสดงผลให้ผู้ใช้) โดยจะใช้ตัวจับเวลาของ OpenMP ระบบที่ใช้ทดสอบจะเป็นคอมพิวเตอร์ส่วนบุคคลแบบธรรมดา (Intel Core i5 4590 at 3.30GHz with DDR3 8GB) ผู้วิจัยได้ใช้ OpenCV ชุด Library สำหรับงานด้านคอมพิวเตอร์วิทัศน์, OpenMP เป็น Library สำหรับการประมวลผลแบบขนาน, และ CLNUI เป็นตัวบันทึกภาพจากกล้องสีและความลึก โดยผู้วิจัยจะทดสอบจำนวนของเลเยอร์และวิธีการจำแนกประเภทของข้อมูลที่แตกต่างกัน ในชุดข้อมูลที่มีทั้งหมดจำนวน 10,720 เฟรม ซึ่งได้แสดงผลของการทดสอบไว้ในตารางที่ 4-8

ตารางที่ 4-8 ผลการทดสอบความเร็วในการรู้จำท่าทางระดับการฟิวชันฟีเจอร์ระดับล่าง

	ระยะเวลาที่ใช้ (ms) / frame									Frame Rate (fps)		
	L=3			L=11			L=19			L=3	L=11	L=19
	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Avg	Avg	Avg
ANN	13.95	17.77	15.08	14.12	20.58	15.18	14.43	20.02	15.85	66.31	65.88	63.09
SVM	14.01	17.40	15.20	14.32	18.97	15.46	14.34	19.64	15.71	65.79	64.68	63.65

โดยที่ L คือจำนวนของเลเยอร์, Min คือ ค่าน้อยสุด, Max คือ ค่ามากสุด, Avg คือ ค่าเฉลี่ย

โดยเฉลี่ยแล้วจะพบว่าเวลาที่ใช้ในการประมวลผลจะใช้เวลาประมาณ 15 มิลลิวินาที (ms) หรือประมวลผลได้ประมาณ 63 เฟรมต่อวินาที (fps) โดยที่ค่าความแตกต่างของจำนวนของเลเยอร์ และวิธีการจำแนกประเภทของข้อมูลแทบไม่มีผลต่อประสิทธิภาพด้านเวลา นอกจากนี้ผู้วิจัยได้ทำการเปรียบเทียบการประมวลผลแบบซีเรียล (Serial Sequence) และการประมวลผลแบบขนาน (Parallel Processing) โดยจะแบ่งจากหน่วยประมวลผลเดียวลงไปในเทรด ซึ่งทำให้การประมวลผลบางอย่างในมุมมองจากกล้องเดี่ยวสามารถทำไปพร้อม ๆ กันได้หลายมุมมองในเวลาเดียวกัน ผลปรากฏว่าการประมวลผลแบบขนานมีความเร็วสูงกว่าแบบซีเรียล 1.5507 เท่า โดยจะใช้เวลาบางอย่างในการเริ่มต้นการประมวลผลแบบขนานไปบ้าง

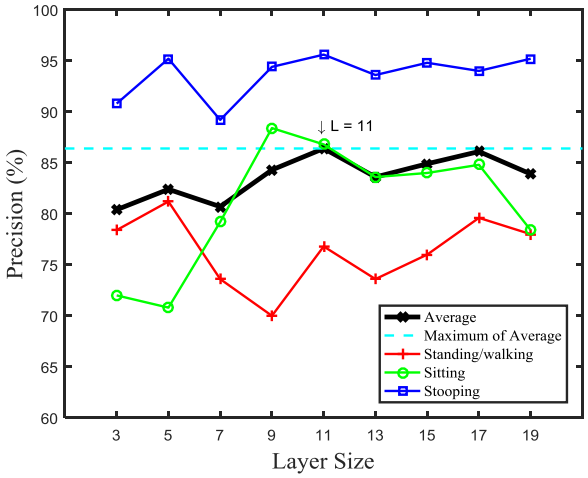
4.3.2 การทดสอบโดยชุดข้อมูล NW-UCLA

ในงานวิจัยนี้จะใช้ชุดข้อมูล NW-UCLA [69] เป็นตัววัดประสิทธิภาพของแบบจำลองฟิวชันแบบเลเยอร์ โดยในชุดข้อมูลนี้มีความใกล้เคียงกับชุดข้อมูลในงานของผู้วิจัย โดยที่ข้อมูลนี้เป็นท่าทางที่ได้จากกล้องสีและความลึก ณ มุมมองที่แตกต่างกัน ชุดข้อมูล NW-UCLA ครอบคลุมถึง 10 ท่าทาง อย่างเช่น ท่ายืนขึ้น, ท่าเดินไปรอบ ๆ , ท่านั่งลง, ท่าก้มไปเก็บของ และท่าอื่น ๆ แต่ยังขาดท่านอน โดยในชุดข้อมูลนี้ผู้วิจัยจะไม่สามารถดึงภาพความลึกที่มีอยู่ในชุดข้อมูลนี้มาใช้งานได้ เนื่องจากค่าความลึกที่แสดงออกมามีค่าแตกต่างกันมาก และไม่สามารถแปลงให้อยู่ในรูปแบบของความลึกในระยะจริง ผู้วิจัยจึงใช้ความเคลื่อนไหวในความลึกที่ถูกเน้นเป็นสีพิเศษมา

ใช้ในการทดสอบ โดยจะทำฟีเจอร์ในรูปแบบของเวกเตอร์ของฟีเจอร์แบบที่ไม่ใช้ความลึกจะสร้างขึ้นมาจากการเรียงต่อกันของค่ามวลสารของมิติจากทุกเลเยอร์เพื่อใช้ในการรู้จำและทดสอบต่อไป

โดยท่าทางที่ผู้วิจัยสนใจจะถูกเลือกออกมาเพื่อทดสอบ โดยท่าทางที่อยู่ระหว่างท่าทางที่ผู้วิจัยสนใจจะไม่ถูกนำมา อย่างเช่น ท่าเดิน/ยืน จะถูกดึงออกมาจากท่ายืนขึ้นและท่าเดินรอบ ๆ, ท่านั่งจะถูกดึงออกมาจากท่านั่งลงและท่ายืนขึ้น, ท่าก้มจะถูกดึงออกมาจากท่าก้มเก็บของมือเดียวและสองมือ

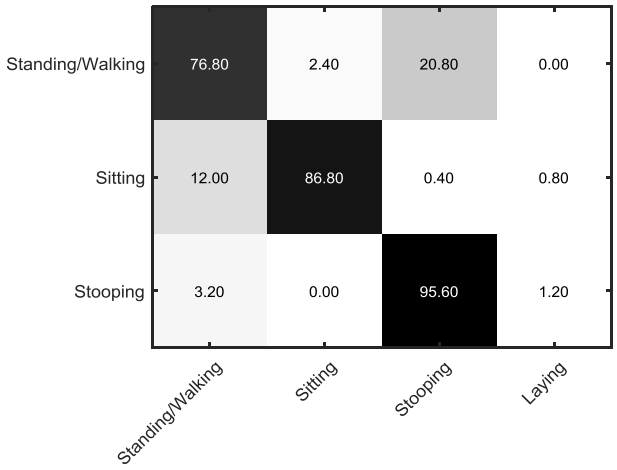
โดยวิธีการของผู้วิจัยที่ใช้ในการรู้จำจะเรียนรู้ข้อมูลมาจากชุดข้อมูล PSU ในฉากห้องนั่งทำงานเพื่อที่จะทดสอบกับข้อมูลในชุดข้อมูล NW-UCLA โดยทุกพารามิเตอร์ที่จะใช้ทดสอบจะมีค่าหรือเหมือนหรือใกล้เคียงกันกับการทดสอบในชุดข้อมูลของ PSU ยกเว้นค่า α ซึ่งจะเท่ากับ 0 เนื่องจากไม่ใช้ภาพความลึกเป็นข้อมูล โดยการทดสอบจะใช้จำนวนของเลเยอร์ที่แตกต่างกันตั้งแต่ L = 3-19 ซึ่งแสดงผลลัพธ์ไว้ในภาพประกอบที่ 4-20



ภาพประกอบที่ 4-20 ความแม่นยำในการรู้จำท่าทางของชุดข้อมูล NW-UCLA ที่มีขนาดของเลเยอร์ที่แตกต่างกัน

จากภาพประกอบที่ 4-20 จะเห็นได้ว่าความแม่นยำสูงสุดโดยเฉลี่ยอยู่ที่ขนาดเลเยอร์ L = 11 (86.40%) ในขณะที่ในชุดข้อมูล PSU ความแม่นยำสูงสุดอยู่ที่ L = 3 โดยประสิทธิภาพสำหรับท่าก้มจะค่อนข้างดีกว่าท่าอื่น ๆ และสูงสุดที่ 95.60% ซึ่งได้แสดงรายละเอียดไว้ในภาพประกอบที่ 4-21 สำหรับท่ายืน/เดิน จะให้ความแม่นยำที่ต่ำที่สุดที่ 76.80% โดยเหตุผลหลัก ๆ อาจจะมาจากมุมของกล้องและช่วงระยะจากกล้องถึงตัวบุคคลมีความแปรปรวนมาก รวมไปถึงมีความแตกต่างกันชัดเจนกับชุดข้อมูล PSU เมื่อเปรียบเทียบวิธีการรู้จำท่าทางของผู้วิจัยกับ

NW-UCLA ปรากฏว่าผู้วิจัยสามารถทำได้ดีกว่าถึง 13 แต้มเปอร์เซ็นต์ จากค่าเฉลี่ย 73.40% และ 86.40% ดังที่ได้แสดงไว้ในตารางที่ 4-9 อย่างไรก็ตามการรู้จำท่าทางของงานวิจัย NW-UCLA จะเน้นหนักไปในส่วนของท่าทางที่มีการเคลื่อนไหวและมีความซับซ้อนที่มากกว่ารวมไปถึงมีจำนวนของท่าทางที่มาใช้ในการรู้จำที่มากกว่า



ภาพประกอบที่ 4-21 Confusion Matrix ของการรู้จำท่าทางในชุดข้อมูล NW-UCLA โดยใช้ L=11, และANN

ตารางที่ 4-9 เปรียบเทียบผลลัพธ์ระหว่างวิธีการรู้จำท่าทางระหว่างงานวิจัย NW-UCLA กับงานวิจัยนี้

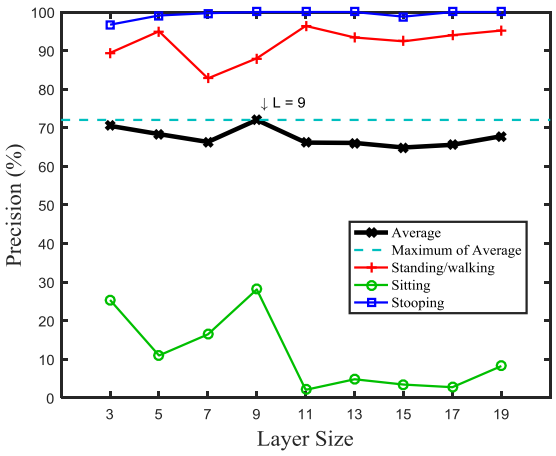
การรู้จำท่าทางของงานวิจัย NW-UCLA [69]	**ความแม่นยำ (%)**	**การรู้จำท่าทางในงานวิจัยนี้**	**ความแม่นยำ (%)**
เดินไปรอบ ๆ	77.60	ยืน/เดิน	76.80
นั่งลง	68.10	นั่ง	86.80
ก้มไปหยิบของมือเดียว	74.50	ก้ม	95.60
เฉลี่ย	**73.40**	**เฉลี่ย**	**86.40**

4.3.3 การทดสอบโดยชุดข้อมูล i3DPost

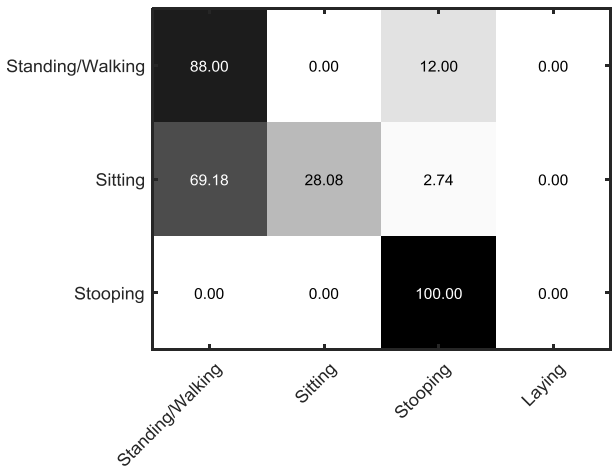
สำหรับการทดสอบในชุดข้อมูล i3DPost [97] โดยในชุดทดสอบได้เลือกเฉพาะท่าทางที่มีความเกี่ยวข้องจากกิจกรรมต่าง ๆ คือ ท่ายืน/เดิน, ท่านั่ง, และท่าก้ม จากกิจกรรม นั่ง-ยืน, เดิน, เดิน-นั่ง, และก้ม โดยข้อมูลภาพจากชุดข้อมูลเป็นภาพสีจึงจะใช้การสร้างเวกเตอร์ของฟีเจอร์แบบไม่ใช้ความลึก ซึ่งใช้ในการรู้จำท่าทางต่อไป

(ก) การทดสอบชุดข้อมูล i3DPost โดยใช้แบบจำลองที่ถูกสอนจากชุดข้อมูล PSU

ในขั้นแรกผู้วิจัยจะทดสอบชุดข้อมูล i3DPost ด้วยแบบจำลองที่ถูกสอนข้อมูลจากชุดข้อมูล PSU จาก 2 มุมมอง ที่ตั้งฉากกันระหว่างกล้องในเลเยอร์ที่แตกต่างกัน จากภาพประกอบที่ 4-22 ได้แสดงถึงผลการทดสอบของชุดข้อมูล i3DPost โดยใช้แบบจำลองที่สอนโดยชุดข้อมูล PSU ซึ่งผลปรากฏว่ามีการเกิดการทำนายผิดสำหรับท่านั่ง ซึ่งได้ผลลัพธ์ความแม่นยำเพียงแค่ 28.08% ที่ L=9 จากการสังเกตการณ์ ปรากฏว่าความผิดพลาดโดยส่วนใหญ่เกิดขึ้นเนื่องจากท่านั่งที่ดูเหมือนท่านั่งยองในอากาศ ซึ่งจะถูกทำนายเป็นท่ายืนเนื่องจากในการสอนใช้ข้อมูลจากชุดข้อมูลของ PSU ซึ่งนั่งอยู่บนม้านั่งหรือเก้าอี้ หรือโซฟา ในทางที่กลับกัน ท่ายืน/เดิน และท่าก้มให้ผลลัพธ์ที่ดีประมาณถึง 96.40% และ 100% โดยที่ L=11 ตามลำดับ



ภาพประกอบที่ 4-22 ความแม่นยำในการรู้จำท่าทางของชุดข้อมูล i3DPost ที่มีขนาดของเลเยอร์ที่แตกต่างกัน โดยใช้แบบจำลองที่ถูกสอนจากชุดข้อมูล PSU



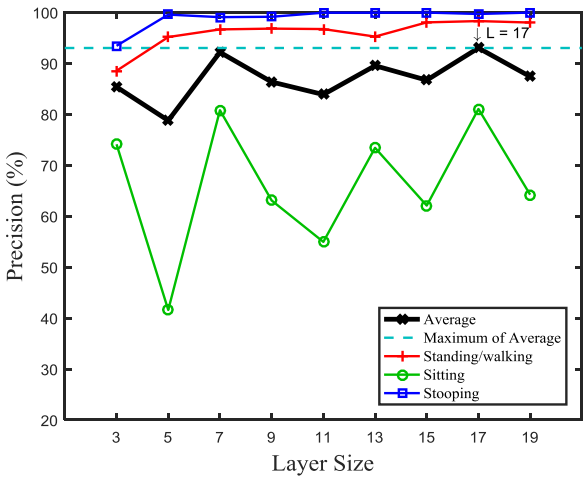
ภาพประกอบที่ 4-23 Confusion Matrix ของการรู้จำท่าทางในชุดข้อมูล i3DPost โดยใช้ L=9, และ ANN โดยใช้แบบจำลองที่ถูกสอนจากชุดข้อมูล PSU

จากภาพประกอบที่ 4-23 ที่แสดงถึง Confusion Matrix โดยเมื่อ L=9 โดยท่านั่งจะตอบผิดเป็นส่วนใหญ่ เป็นท่ายืนและเดิน (69.18% of Cases) และท่ายืน/เดิน จะตอบผิดไปเป็นท่าก้ม (12.00% of Cases)

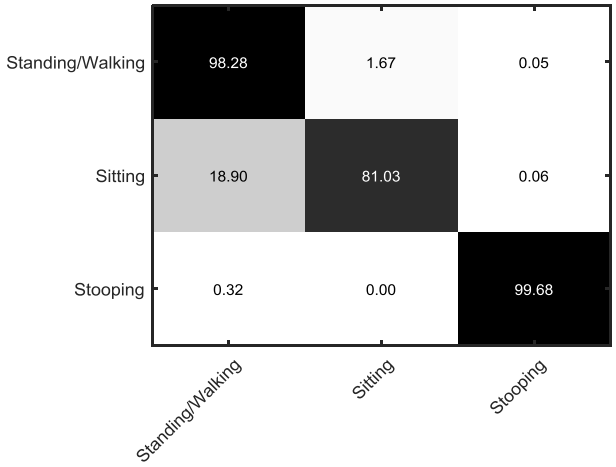
(ข) การทดสอบในชุดข้อมูล i3DPost โดยการสอนข้อมูลใหม่

เนื่องจากการทดสอบในชุดข้อมูล i3DPost โดยใช้แบบจำลองที่ถูกสอนโดยชุดข้อมูล PSU ให้ผลลัพธ์ของการแยกแยะผิดในท่านั่ง ดังนั้น ผู้วิจัยจึงได้ทำการทดสอบด้วยแบบจำลองของผู้วิจัย ซึ่งถูกสอนและประเมินโดยใช้เฉพาะชุดข้อมูลของ i3DPost เท่านั้น โดยชุดข้อมูลชุดแรกของ i3DPost จะถูกใช้สำหรับการทดสอบและข้อมูลชุดที่สองจะถูกใช้เพื่อการสอนข้อมูล โดยในการทดสอบขั้นต้นจะเริ่มทดสอบจากการทดสอบโดยใช้ 2 มุมมองก่อน

จากภาพประกอบที่ 4-24 และ 4-25 แสดงให้เห็นถึงผลลัพธ์ในขนาดของเลเยอร์ที่แตกต่างกัน จาก 2 มุมมอง โดยการใช้เลเยอร์ที่มีขนาดเท่ากับ 17 ให้ผลลัพธ์ความแม่นยำที่สูงสุดถึง 93.00% โดยเฉลี่ย (98.28%, 81.03% และ 99.68% สำหรับท่ายืน/เดิน, นั่งและก้ม) โดยส่วนใหญ่แล้ว ท่ายืน/เดิน และก้ม ให้ผลลัพธ์ค่อนข้างดีมาก ซึ่งสูงกว่า 90% ยกเว้น L=3 อย่างไรก็ตาม ค่าความแม่นยำที่ดีที่สุดสำหรับท่านั่ง ได้ผลเพียงแค่ 81.03% โดยที่จะทำนายผิดไปเป็นท่ายืน/เดิน (18.90%) และมีความแม่นยำต่ำสุดอยู่ที่ 41.59% เมื่อ L=5 ซึ่งสังเกตได้ว่าท่านั่งยองยังคงมีผลต่อประสิทธิภาพในการทำนาย

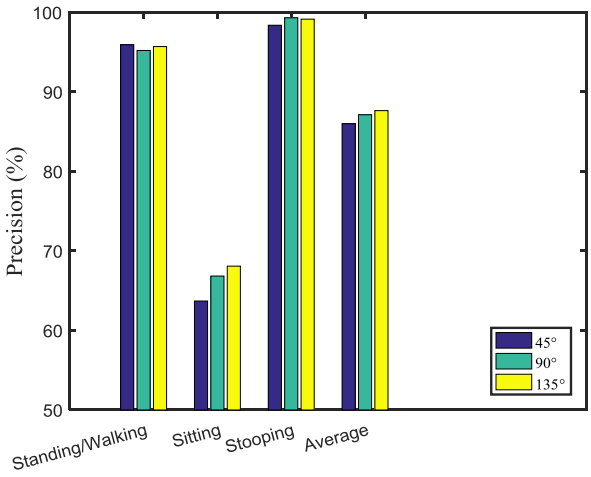


ภาพประกอบที่ 4-24 ความแม่นยำในการรู้จำท่าทางของชุดข้อมูล i3DPost ที่มีขนาดของเลเยอร์ที่แตกต่างกัน โดยใช้แบบจำลองที่ถูกสอนใหม่จากชุดข้อมูล i3DPost



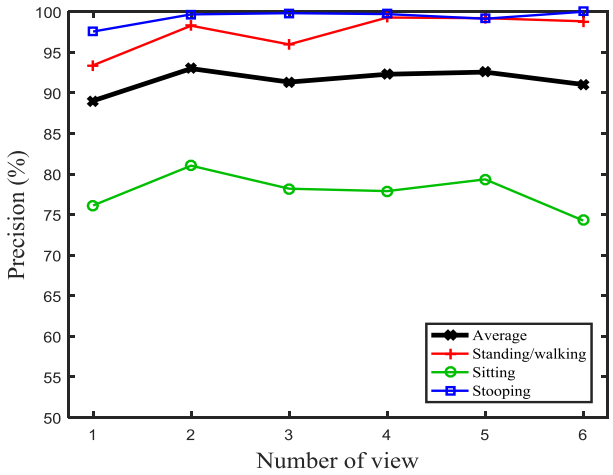
ภาพประกอบที่ 4-25 Confusion Matrix ของการรู้จำท่าทางในชุดข้อมูล i3DPost โดยใช้ L=17, และ ANN โดยใช้แบบจำลองที่ถูกสอนใหม่จากชุดข้อมูล i3DPost

ในการทดสอบ 2 มุมมอง ผู้วิจัยได้จับคู่มุมมองระหว่างกล้องที่แตกต่างกัน เช่น 45°, 90° และ135° ภาพประกอบที่ 4-26 แสดงให้เห็นว่าที่มุมเท่ากับ 135° จะมีความแม่นยำสูงที่สุด และความแม่นยำต่ำที่สุดในมุมที่เท่ากับ 45° โดยส่วนใหญ่แล้วมุมที่แคบลงจะให้ผลลัพธ์ที่ด้อยกว่า อย่างไรก็ตาม สำหรับในท่านั่ง มุมที่แคบอาจจะลดความแม่นยำลงเป็นอย่างมาก



ภาพประกอบที่ 4-26 การเปรียบเทียบผลลัพธ์ความแม่นยำในชุดข้อมูล i3DPost จากการทำมุมกันของกล้องที่องศาแตกต่างกัน โดยใช้ L=17, และANN โดยใช้แบบจำลองที่ถูกสอนใหม่จากชุดข้อมูล i3DPost

นอกจากนี้ ผู้วิจัยยังได้ทำการทดสอบโดยใช้จำนวนของมุมมองที่แตกต่างกัน นอกจากการทดสอบ 2 มุม โดยจะทดสอบตั้งแต่จำนวนมุมที่เท่ากับ 1 ถึง 6 ตามลำดับ เพื่อที่จะประเมินแบบจำลองของผู้วิจัย ซึ่งผลลัพธ์ดังภาพประกอบที่ 4-27



ภาพประกอบที่ 4-27 ความแม่นยำในการรู้จำท่าทางของชุดข้อมูล i3DPost ที่มีจำนวนของมุมมองที่แตกต่างกัน โดยใช้แบบจำลองที่ถูกสอนใหม่จากชุดข้อมูล i3DPost

ภาพประกอบที่ 4-27 แสดงถึงความแม่นยำตามจำนวนของมุมมองที่แตกต่างกัน กราฟนี้ได้รายงานผลความแม่นยำสูงสุดในแต่ละมุมมองจาก 1 ถึง 6 ซึ่งมีค่าความแม่นยำดังนี้ 89.03%, 93.00%, 91.33%, 92.30%, 92.56% และ 91.03% ที่ L=7, L=17, L=17, L=7, และ L=13, ตามลำดับ ผู้วิจัยได้ตั้งข้อสังเกตว่าค่าความแม่นยำสูงที่สุดอยู่ ณ จำนวนของมุมมองที่เท่ากับ 2 โดยทั่วไปแล้ว ประสิทธิภาพในการทำนายจะลดลงเมื่อจำนวนของมุมมองเพิ่มขึ้น

ยกเว้นในท่านั่ง นอกจากนี้ จำนวนของเลเยอร์ที่ให้ค่าความแม่นยำที่สูงที่สุดจะลดลงเมื่อจำนวนของมุมมองเพิ่มขึ้น

ผู้วิจัยได้เปรียบเทียบผลลัพธ์จากวิธีการของผู้วิจัยกับงานอื่นที่ใกล้เคียงกันและใช้ชุดข้อมูล i3DPost ในการทดสอบ [56] ซึ่งเป็นงานที่ใช้ฟีเจอร์แผนผังต้นแบบของท่า (Posture Prototype Map) ซึ่งจัดข้อมูลด้วย Self-organizing Map เพื่อรู้จำท่าทางในมุมมองเดี่ยวร่วมกับ Traditional Neural Network (TNN) ซึ่งตารางที่ 4-10 ได้แสดงถึงผลลัพธ์ของงานทั้ง 2 โดยความแม่นยำสูงสุดจากวิธีการของผู้วิจัย และงานที่ใกล้เคียงคือ 99.68% และ 100% สำหรับท่าก้ม ในขณะเดียวกันความแม่นยำที่ต่ำสุดอยู่ในท่านั่งวิธีการของผู้วิจัย และงานที่ใกล้เคียงที่ 81.03% และ 87.00% ตามลำดับ อย่างไรก็ตาม ในท่ายืน/เดิน งานของผู้วิจัยให้ผลลัพธ์ที่ค่อนข้างดีกว่าเล็กน้อย โดยสรุปแล้วทั้งสองงานมีผลลัพธ์ที่ใกล้เคียงกัน ซึ่งงานที่ใกล้เคียงกันสามารถทำได้ออกมาดีกว่าเล็กน้อย

ตารางที่ 4-10 เปรียบเทียบผลลัพธ์ระหว่างวิธีการรู้จำท่าทางระหว่างงานวิจัยที่ใกล้เคียงกันซึ่งใช้ชุดข้อมูล i3DPost [56] กับงานวิจัยนี้

การรู้จำท่าทางในงานวิจัยที่ใกล้เคียงกัน [56]	**ความแม่นยำ (%)**	**การรู้จำท่าทางในงานวิจัยนี้**	**ความแม่นยำ (%)**
เดินไปรอบ ๆ	95.00	ยืน/เดิน	98.28
นั่งลง	87.00	นั่ง	81.03
ก้ม (Bend)	100.0	ก้ม (Stooping)	99.68
ค่าเฉลี่ย	**94.00**	**ค่าเฉลี่ย**	**93.00**

4.3.4 วิเคราะห์ผลการทดสอบการรู้จำท่าทางระดับการฟิวชันฟีเจอร์ในระดับล่าง

จากผลการทดสอบที่ได้เปรียบเทียบประสิทธิภาพของการรู้จำจากมุมมองเดี่ยวและหลายมุมมองจากชุดข้อมูล PSU และ i3DPost จะเห็นได้ว่าการใช้หลายมุมมองจะให้ความแม่นยำที่มากกว่ามุมมองเดียว ซึ่งค่าเฉลี่ยของความแม่นยำในมุมมองเดี่ยวเท่ากับ 92.50% และ 90.63% ขณะที่สองมุมมองมีค่าเฉลี่ยของความแม่นยำอยู่ที่ 95.32% ในชุดข้อมูล PSU ในขณะที่ค่าเฉลี่ยของความแม่นยำในชุดข้อมูล i3DPost เป็น 89.03%, 93.00%, 91.33%, 92.30%, 92.56% (เรียงจากจำนวนของมุมมองที่ 1-6) ซึ่งก็มีแนวโน้มที่เพิ่มขึ้นจากมุมมองเดี่ยว ซึ่งสรุปได้ว่าจำนวนของมุมมองเพียงแค่ 2 หรือ 3 จากมุมมองที่แตกต่างกันก็เพียงพอแล้วสำหรับการรู้จำท่าทางในแบบจำลองของผู้วิจัย โดยที่การมีมุมมองที่มากขึ้นอาจจะมีข้อมูลที่มากเกินจำเป็น และยังเป็นการสะสมสิ่งรบกวนได้

ซึ่งผู้วิจัยรายงานการทดสอบที่ใช้ข้อมูลที่สอนและทดสอบต่างชุดข้อมูลต่างกัน เพื่อทดสอบความทนทานของแบบจำลองในการรู้จำท่าทางระดับการฟิวชันฟีเจอร์ในระดับล่าง ได้แก่ (1) ทดสอบกับชุดข้อมูล PSU ในฉากห้องนั่งเล่น ซึ่งสอนข้อมูลฉากห้องทำงาน มีความแม่นยำที่สูงที่สุดจากค่าเฉลี่ย 95.32% (2) ทดสอบกับชุดข้อมูล PSU ซึ่งสอนข้อมูลโดย NW-UCLA มีความแม่นยำที่สูงที่สุดจากค่าเฉลี่ยที่ 93.44% ซึ่งได้ผลลัพธ์ค่อนข้างดีในระดับหนึ่ง, (3) ทดสอบกับชุดข้อมูล NW-UCLA ซึ่งสอนข้อมูลโดย PSU มีความแม่นยำโดยเฉลี่ยอยู่เพียงแค่ 86.40% โดยที่ชุดข้อมูล NW-UCLA มีการวางมุมของกล้องและช่วงระยะจากกล้องถึงตัวบุคคลมีความแปรปรวนมาก รวมไปถึงมีความแตกต่างกันชัดเจนกับชุดข้อมูล PSU, (4) ทดสอบกับชุดข้อมูล i3DPost ซึ่งสอนข้อมูลโดย PSU มีความแม่นยำโดยเฉลี่ยอยู่ที่ 72.03% เนื่องจากท่านั่งของชุดข้อมูล i3DPost เป็นท่านั่งยองในอากาศจึงทำให้ท่านั่งถูกทำนายเป็นท่ายืน ในทางที่กลับกัน ท่ายืน/เดิน และท่าก้มให้ผลลัพธ์ที่ค่อนข้างดีถึง 96.40% และ 100% ซึ่งจากผลการทดสอบที่ใช้ข้อมูลที่สอนและทดสอบต่างชุดข้อมูลกัน ผู้วิจัยให้ความเห็นว่าแบบจำลองในการรู้จำท่าทางนี้มีความทนทานต่อการเปลี่ยนแปลงในระดับหนึ่ง แต่มุมกล้องที่ติดตั้งต่างกันจากข้อมูลมาก ๆ รวมไปถึงท่าทางที่นอกเหนือจากการสอนข้อมูล ก็มีผลต่อประสิทธิภาพของระบบบ้าง นอกจากนี้ผู้วิจัยยังได้ทดสอบการทำมุมกันของกล้องที่องศาแตกต่างกันจากการสอนข้อมูลที่กล้องตั้งฉากกัน ซึ่งผลปรากฏว่าระบบก็ยังสามารถรู้จำได้ดีอยู่ แต่ความแม่นยำในบางท่าทางจะลดลงบ้าง แต่เมื่อมองโดยภาพรวมจะลดลงเล็กน้อย

4.4 การทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง

การทดสอบจะใช้คอมพิวเตอร์ที่มีหน่วยประมวลผล CPU Intel Core i5 4590 ที่ความถี่ 3.30 GHz โดยใช้ OpenCV เป็น Library สำหรับประมวลผลด้าน Computer Vision และ CLNUI Library สำหรับติดต่อกับกล้อง Kinect และรับส่งข้อมูลภาพและข้อมูลความลึก โดยได้แบ่งการประมวลผลในมุมมองเดี่ยวให้อยู่ใน Thread และประมวลผลพร้อม ๆกันแบบขนานโดยใช้ OpenMP Library โดยผลการทดสอบแสดงจะแสดงถึงค่า True Positive (TP), False Negative (FN), False Positive (FP) และสรุปเป็นค่า Precision และ Recall

การทดสอบจะถูกแบ่งเป็นสองประเภทโดยประเภทแรกจะเป็นการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละหนึ่งคน (Single Tracking & Re-identification) และประเภทที่สองจะเป็นการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละสองคน (Multiple Tracking & Re-identification)

4.4.1 ผลการทดสอบการติดตามและจดจำโดยเข้าไปในระบบครั้งละหนึ่งคน

โดยจะมีข้อมูลเป็น Video ชุดแรกเข้าไปให้ระบบติดตามและจดจำ ต่อมาจะเป็น Video ทดสอบ โดยจะทดสอบพิเศษเฉพาะตามท่าทางต่าง ๆในหลาย ๆมุมมอง เนื่องจากการเปลี่ยนท่าทางและมุมมอง โดยท่าทางที่แตกต่างกันจะทำให้มีลักษณะภายนอกที่แตกต่างกันและมุมมองจะทำให้แสงสว่างเปลี่ยนซึ่งความเข้มของสีที่ใช้ในการติดตามและจดจำบุคคลก็จะเปลี่ยนไปด้วย โดยแต่ละชุดของข้อมูลจะเริ่มจากการไปยืนกลางห้องให้ระบบสามารถติดตามและจดจำได้ก่อนจึงเริ่มเปลี่ยนท่าทาง โดยระบบจะนับความถูกต้องเฉพาะท่าทางนั้น ๆที่ทำการทดสอบ

โดยข้อมูลที่ใช้ในการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละคน (Single Tracking & Re-identification) จะมีทั้งหมด 3 Datasets โดยแต่ละ Dataset จะมีบุคคลที่เข้าทดสอบ 5 บุคคล โดยแต่ละบุคคลมีลักษณะสีเสื้อที่แตกต่างกันชัดเจน โดยทดสอบทั้งหมด 4 ท่าทาง คือ ยืนและเดิน (Standing & Walking), นั่ง (Sitting), ก้ม (Bending) และนอน (Laying) แต่ละ Dataset จะวัดผลการติดตามและจดจำบุคคลในระดับเฟรมเพื่อความละเอียดในการสรุปผลการทดลอง โดยผลการทดสอบได้แสดงไว้ตามตารางที่ 4-11 สำหรับ Dataset ชุดที่ 1, ตามตารางที่ 4-12 สำหรับ Dataset ชุดที่ 2 และตามตารางที่ 4-13 สำหรับ Dataset ชุดที่ 3

ตารางที่ 4-11 ผลการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละหนึ่งคนสำหรับ Dataset # 1

Action	True Positive	False Positive	False Negative	All	Precision	Recall
ยืน / เดิน	679	174	7	860	79.60%	98.98%
นั่ง	570	0	0	570	100%	100%
ก้ม	614	0	0	614	100%	100%
นอน	380	110	0	490	77.55%	100%

ตารางที่ 4-12 ผลการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละหนึ่งคนสำหรับ Dataset # 2

ท่าทาง	True Positive	False Positive	False Negative	All	Precision	Recall
ยืน / เดิน	799	0	0	799	100%	100%
นั่ง	495	0	0	495	100%	100%
ก้ม	482	0	0	482	100%	100%
นอน	452	61	0	513	88.11%	100%

ตารางที่ 4-13 ผลการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละหนึ่งคนสำหรับ Dataset # 3

ท่าทาง	True Positive	False Positive	False Negative	All	Precision	Recall
ยืน / เดิน	649	0	0	649	100%	100%
นั่ง	482	0	1	483	100%	99.79%
ก้ม	401	53	16	470	88.33%	96.16%
นอน	445	105	0	550	80.91%	100%

โดยการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละหนึ่งคนจะมีความแม่นยำสูงในทุกท่าทาง แต่ในท่านอน (Laying) จะมีค่าน้อยกว่าท่าทางอื่นเพราะท่านี้จะเปลี่ยนทิศทางและลักษณะของตัววัตถุมากกว่าท่าทางอื่น ๆ รวมไปถึงการตรวจจับไม่ได้ทุกส่วนของตัวบุคคล เนื่องจาก Sensor ความลึกไม่สามารถแยกบุคคลออกจากพื้นหลังได้ดีเท่าที่ควร เนื่องจากบุคคลใน

ท่านอนจะมีความลึกใกล้เคียงกับพื้นหลัง โดยจะมีค่าเฉลี่ย (Precision) การติดตามและจดจำในท่ายืนและเดิน (Standing & Walking) 93.20%, ท่านั่ง (Sitting) 100%, ท่าก้ม (Bending) 96.10% และท่านอน 82.19%

สำหรับการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละหนึ่งคนได้บันทึกไว้ใน Youtube.com
ซึ่งสามารถเข้าถึงได้ที่

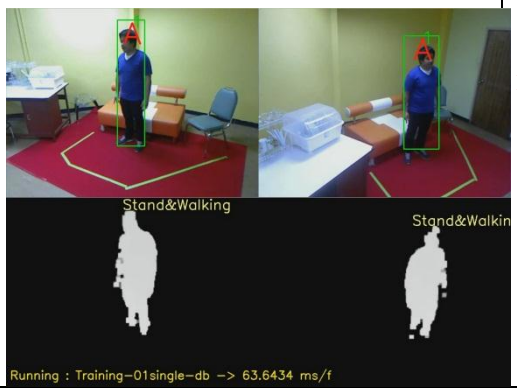
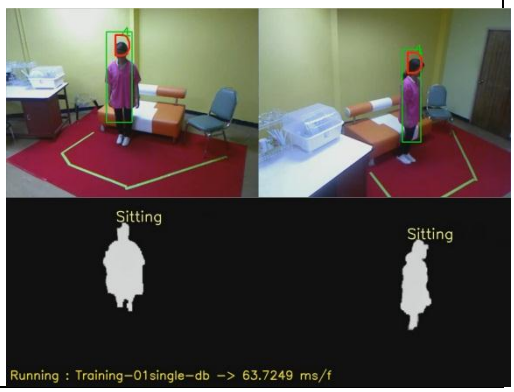
Dataset#1 https://www.youtube.com/watch?v=SFuyGYVgaIE

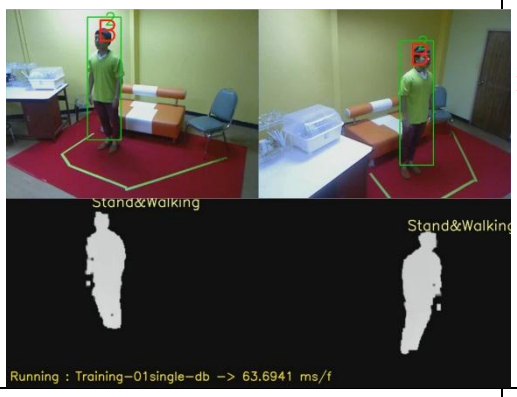
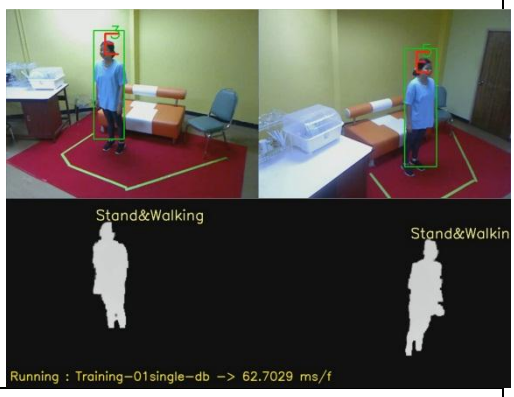
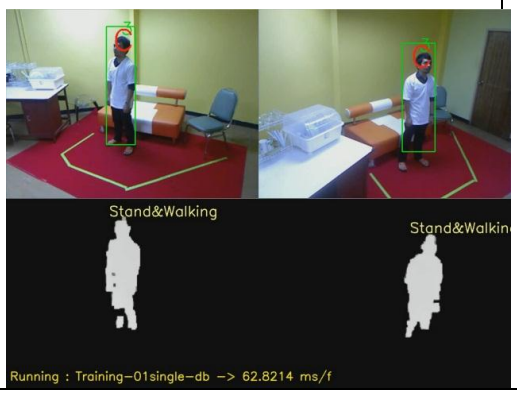
Dataset#2 https://www.youtube.com/watch?v=sX33H0F5Yy8

Dataset#3 https://www.youtube.com/watch?v=WJYOF1b-FOY

ตัวอย่างบุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก ในชุดทดสอบชุดที่ 1 ได้แสดงไว้ในตารางที่ 4-14 และตัวอย่างการทำงานของระบบติดตามและรู้จำในชุดทดสอบชุดที่ 1 ได้แสดงไว้ในตารางที่ 4-15

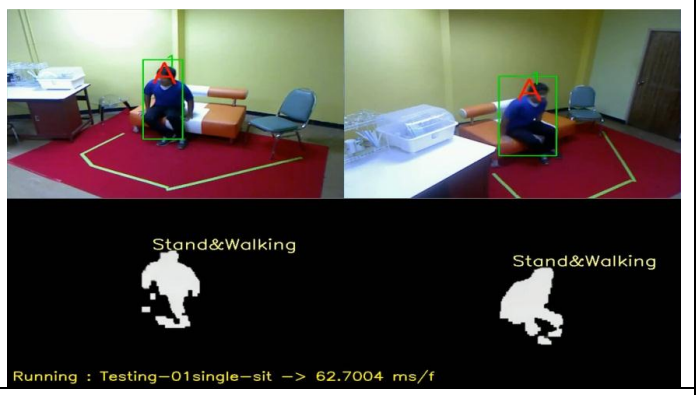
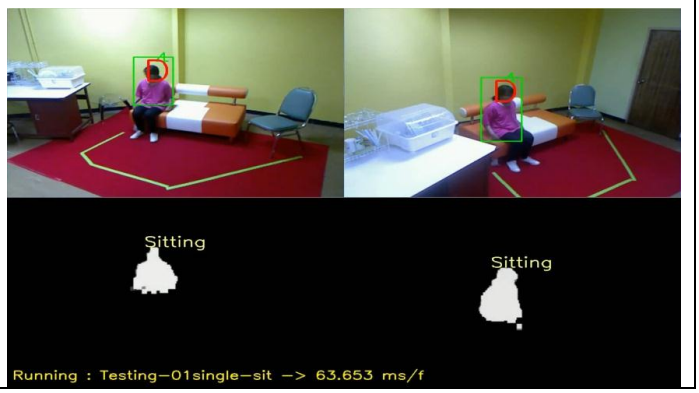
ตารางที่ 4-14 ตัวอย่างบุคคลที่เข้าทดสอบและ Global ID สำหรับ Dataset # 1

ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก	ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก
#1 A		#4 D	

ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก	ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก
#2 B		#5 E	
#3 C		-	-

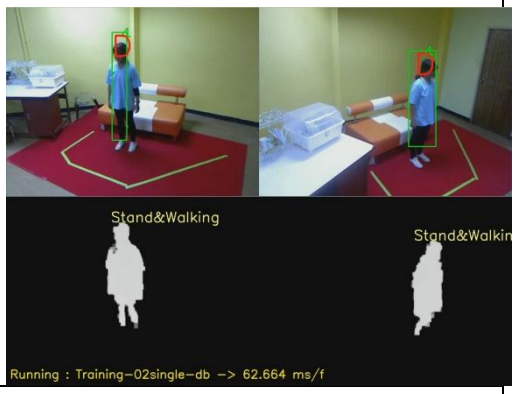
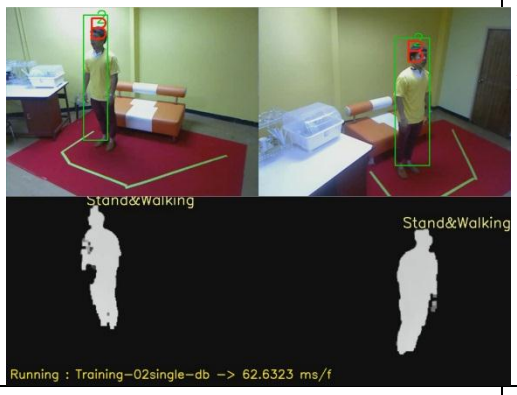
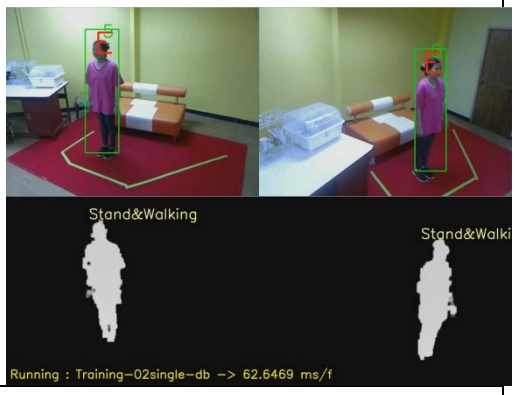
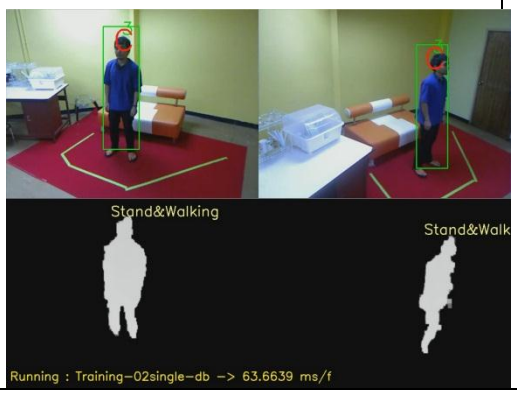
ตารางที่ 4-15 ตัวอย่างการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละหนึ่งคนสำหรับ Dataset # 1

ตัวอย่างที่	ตัวอย่างเฟรมของการติดตามและจดจำ
1	

ตัวอย่างที่	ตัวอย่างเฟรมของการติดตามและจดจำ
2	
3	
4	

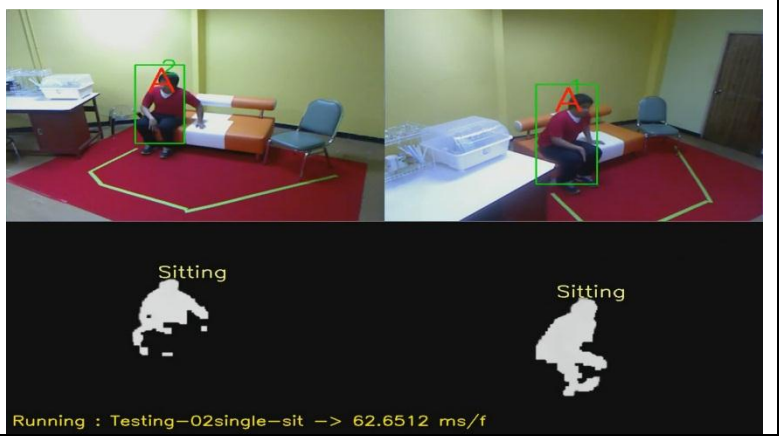
ตัวอย่างบุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก ในชุดทดสอบชุดที่ 2 ได้แสดงไว้ในตารางที่ 4-16 และตัวอย่างการทำงานของระบบติดตามและรู้จำในชุดทดสอบชุดที่ 2 ได้แสดงไว้ในตารางที่ 4-17

ตารางที่ 4-16 ตัวอย่างบุคคลที่เข้าทดสอบและ Global ID สำหรับ Dataset # 2

ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก	ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก
#1 A		#4 D	
#2 B		#5 E	
#3 C		-	-

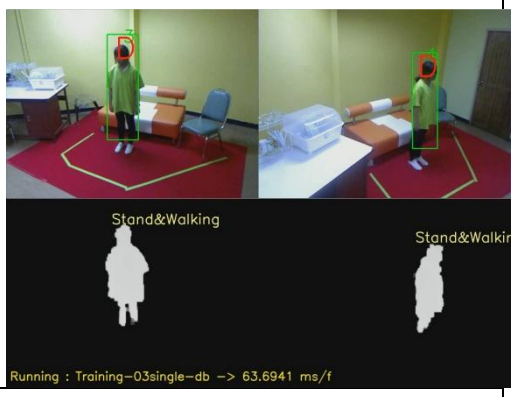
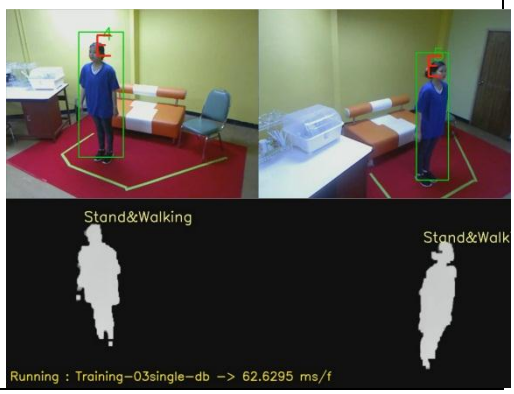
ตารางที่ 4-17 ตัวอย่างการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละหนึ่งคนสำหรับ Dataset # 2

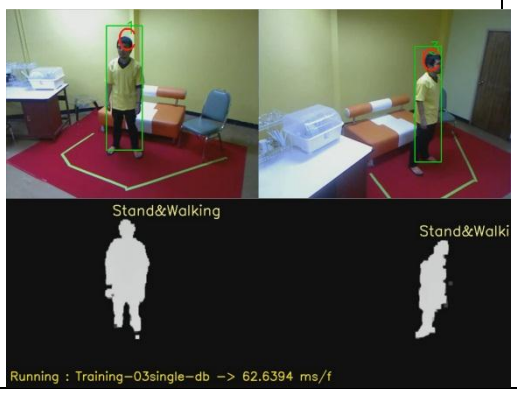
ตัวอย่างที่	ตัวอย่างเฟรมของการติดตามและจดจำ
1	
2	
3	

ตัวอย่างที่	ตัวอย่างเฟรมของการติดตามและจดจำ
4	

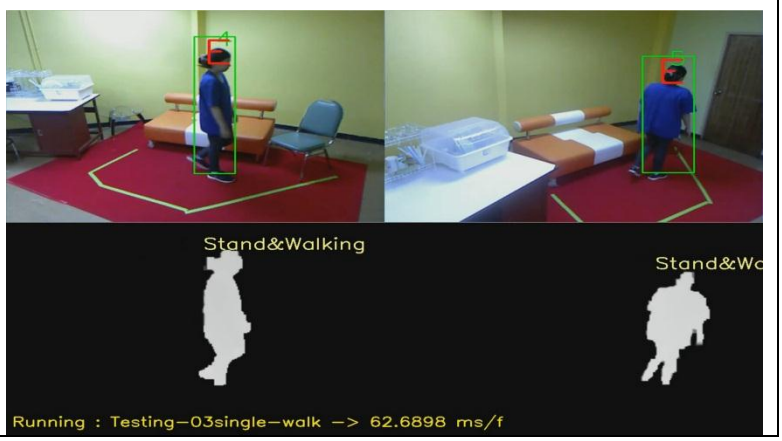
ตัวอย่างบุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก ในชุดทดสอบชุดที่ 3 ได้แสดงไว้ในตารางที่ 4-18 และตัวอย่างการทำงานของระบบติดตามและรู้จำในชุดทดสอบชุดที่ 3 ได้แสดงไว้ในตารางที่ 4-19

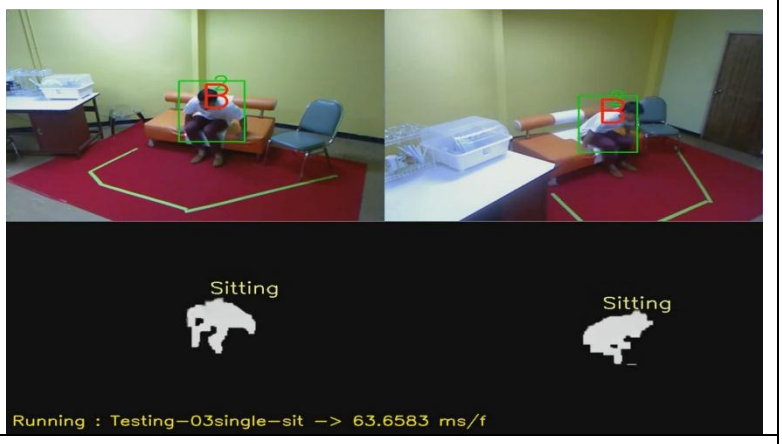
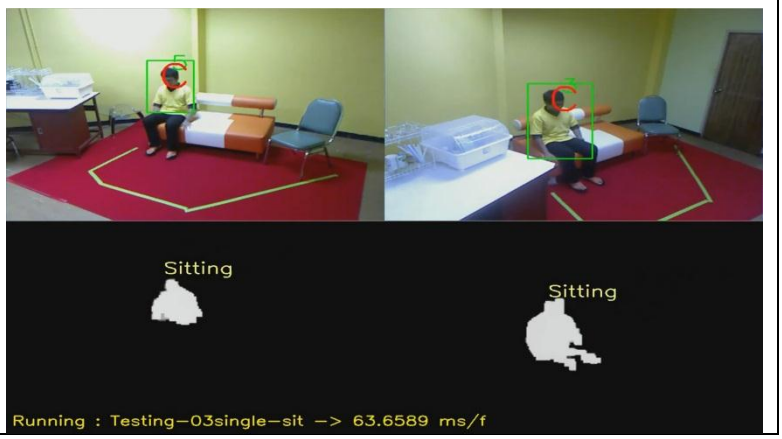
ตารางที่ 4-18 ตัวอย่างบุคคลที่เข้าทดสอบและ Global ID สำหรับ Dataset # 3

ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก	ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก
#1 A		#4 D	
#2 B		#5 E	

ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก	ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก
#3 C		-	-

ตารางที่ 4-19 ตัวอย่างการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละหนึ่งคนสำหรับ Dataset # 3

ตัวอย่างที่	ตัวอย่างเฟรมของการติดตามและจดจำ
1	
2	

ตัวอย่างที่	ตัวอย่างเฟรมของการติดตามและจดจำ
3	
4	

4.4.2 ผลการทดสอบการติดตามและจดจำโดยเข้าไปในระบบครั้งละสองคน

โดยจะมีข้อมูลที่เป็น Video ชุดแรกป้อนเข้าไปให้ระบบติดตามและจดจำ โดยเข้าไปครั้งละหนึ่งคนเช่นเดียวกับการทดสอบประเภทแรก ต่อมาจะเป็น Video ทดสอบโดยเข้าไปที่ละหนึ่งคนจนครบสองคน โดยจะทดสอบพิเศษเฉพาะตามท่าทางต่าง ๆในหลาย ๆมุมมองเช่นเดียวกับการทดสอบประเภทแรก โดยแต่ละชุดของข้อมูลจะเริ่มจากการไปยืนตามตำแหน่งที่จะไม่มีการบดบังเพื่อให้ระบบสามารถติดตามและจดจำได้ก่อนจึงเริ่มเปลี่ยนท่าทางในแต่ละการทดสอบ โดยระบบจะนับความถูกต้องเฉพาะท่าทาง ๆนั้น ๆที่ทำการทดสอบ โดยจะมีการเดินเพื่อสลับตำแหน่งที่ทดสอบทั้งหมด 3 ครั้ง

โดยการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละสองคนจะมีทั้งหมด 2 Datasets โดยแต่ละ Dataset จะมีบุคคลที่เข้าทดสอบ 4 บุคคล โดยมีลักษณะสีเสื้อที่แตกต่างกันชัดเจน ซึ่งในแต่ละ Dataset จะวัดผลการติดตามและจดจำบุคคลในระดับเฟรมเพื่อความละเอียดในการสรุปผลการทดลอง โดยทดสอบทั้งหมด 3 ท่าทาง คือ ยืนหรือเดิน, นั่ง, และก้ม เนื่องจากท่านอนในบริเวณที่ไม่บดบังกันต้องในพื้นที่กว้างที่ทำผู้ทดลองต้องอยู่ไกลจากกล้อง เป็นผลให้ไม่สามารถ

ตรวจจับได้เนื่องจากเซ็นเซอร์ความลึกไม่สามารถแยกบุคคลออกจากพื้นหลังได้ โดยผลการทดสอบได้แสดงไว้ตามตารางที่ 4-20 สำหรับ Dataset ชุดที่ 1 และตารางที่ 4-21 สำหรับ Dataset ชุดที่ 2

ตารางที่ 4-20 ผลการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละสองคนสำหรับ Dataset # 1

ท่าทาง	True Positive	False Positive	False Negative	All	Precision	Recall
ยืน / เดิน	728	214	0	942	77.28%	100%
นั่ง	875	179	0	1054	83.02%	100%
ก้ม	1112	370	42	1524	75.03%	96.36%

ตารางที่ 4-21 ผลการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละสองคนสำหรับ Dataset # 2

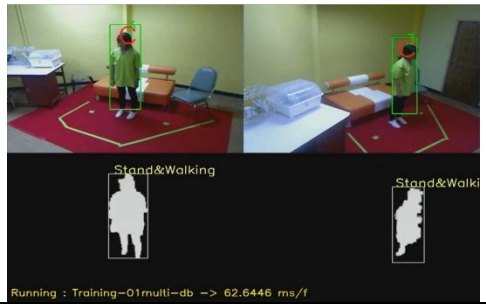
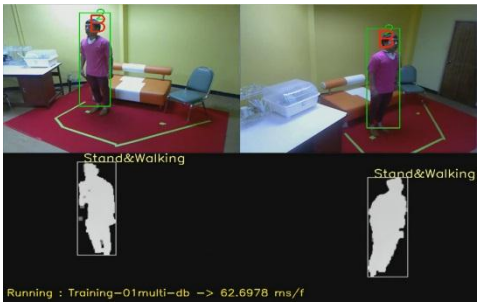
ท่าทาง	True Positive	False Positive	False Negative	All	Precision	Recall
ยืน / เดิน	1105	48	9	1162	95.84%	99.19%
นั่ง	905	121	0	1026	88.21%	100%
ก้ม	952	110	0	1062	89.64%	100%

จากการทดสอบทั้งสอง Dataset จะมีค่าเฉลี่ยความแม่นยำการติดตามและจดจำในท่ายืนและเดิน 88.56%, ท่านั่ง 85.61% และท่าก้ม 82.34%

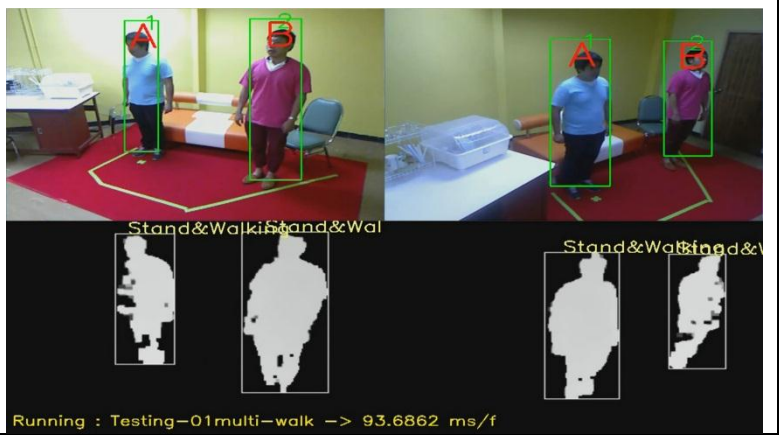
สำหรับการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละสองคนได้บันทึกไว้ใน Youtube.com
ซึ่งสามารถเข้าถึงได้ที่
Dataset#1 https://www.youtube.com/watch?v=ij7lQa2S5F0
Dataset#2 https://www.youtube.com/watch?v=i7g985ekTNc

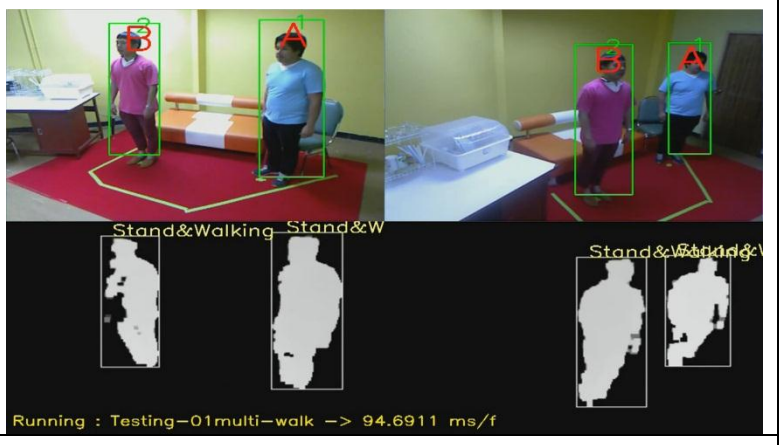
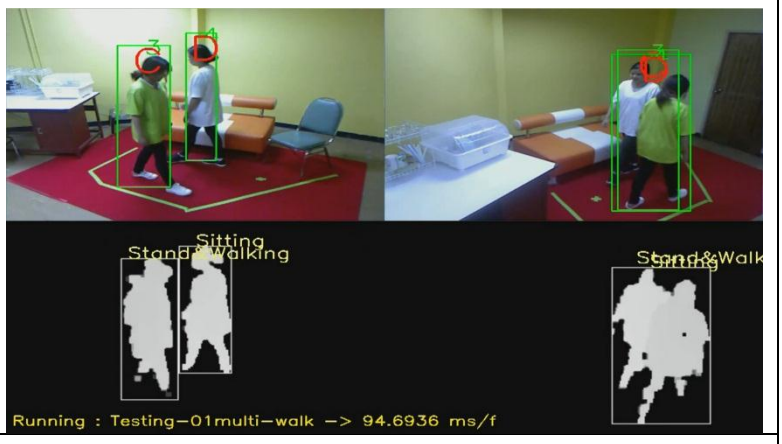
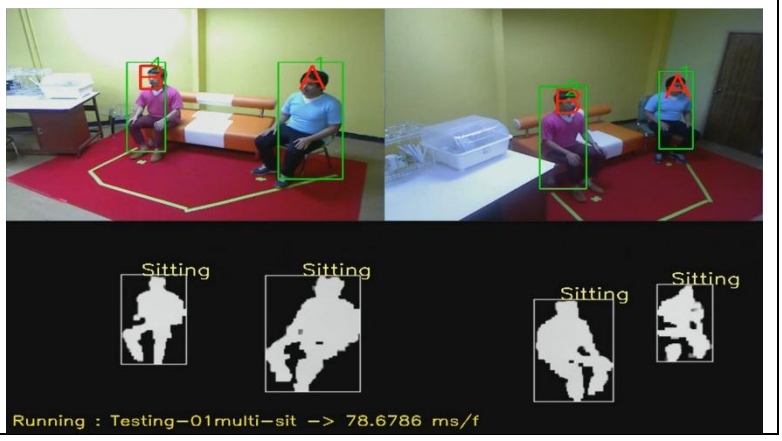
ตัวอย่างบุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก ในชุดทดสอบครั้งละสองคนชุดที่ 1 ได้แสดงไว้ในตารางที่ 4-22 และตัวอย่างการทำงานของระบบติดตามและรู้จำครั้งละสองคนในชุดทดสอบชุดที่ 1 ได้แสดงไว้ในตารางที่ 4-23

ตารางที่ 4-22 ตัวอย่างบุคคลที่เข้าทดสอบครั้งละสองคนและ Global ID สำหรับ Dataset # 1

ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก	ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก
#1 A		#3 C	
#2 B		#4 D	

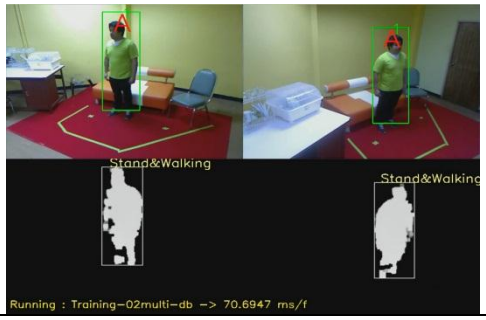
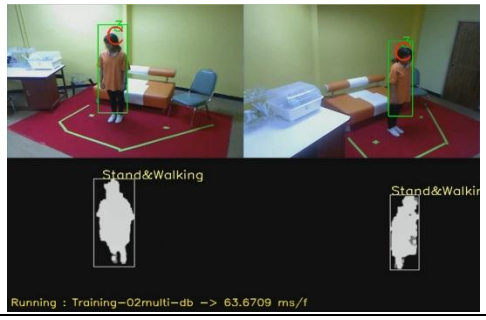
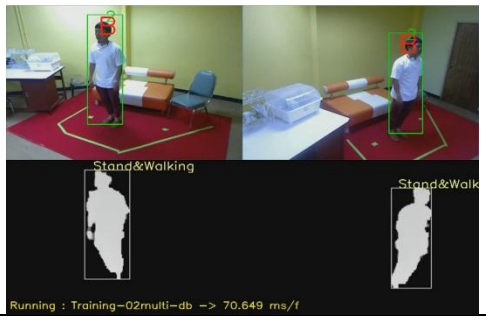
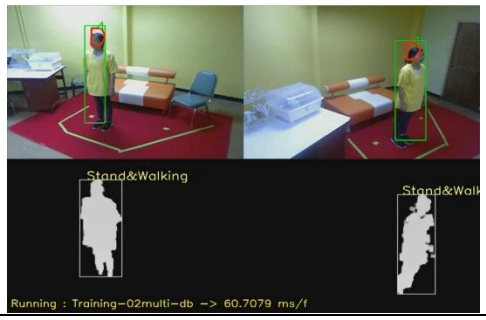
ตารางที่ 4-23 ตัวอย่างการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละสองคนสำหรับ Dataset # 1

ตัวอย่างที่	ตัวอย่างเฟรมของการติดตามและจดจำ
1	

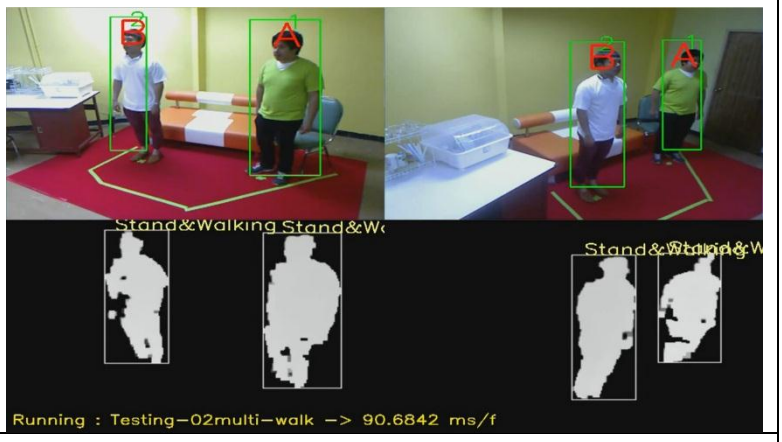
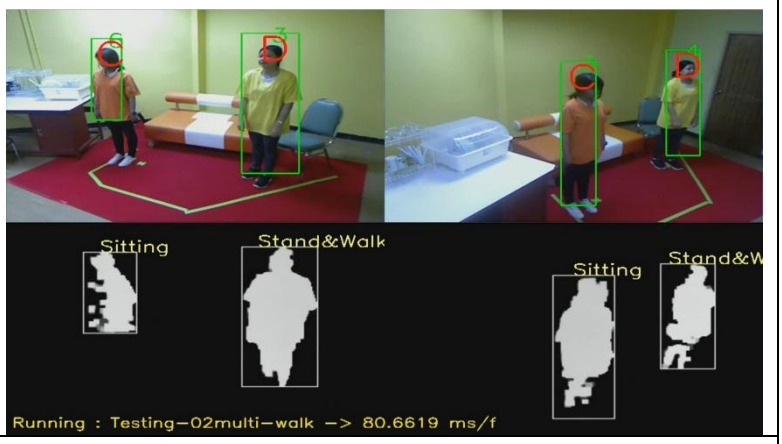
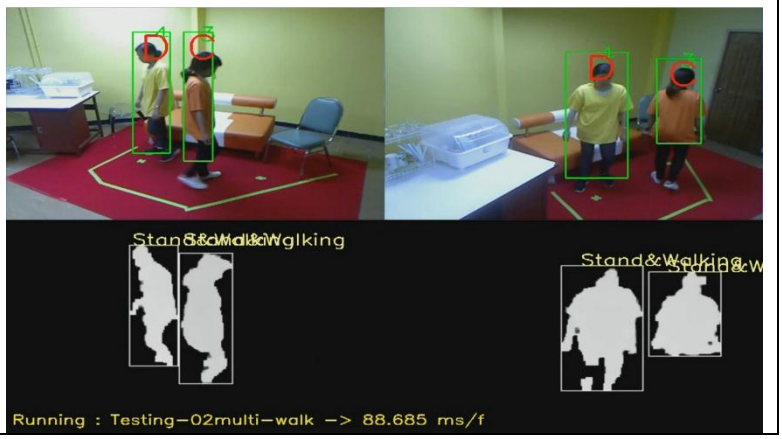
ตัวอย่างที่	ตัวอย่างเฟรมของการติดตามและจดจำ
2	
3	
4	

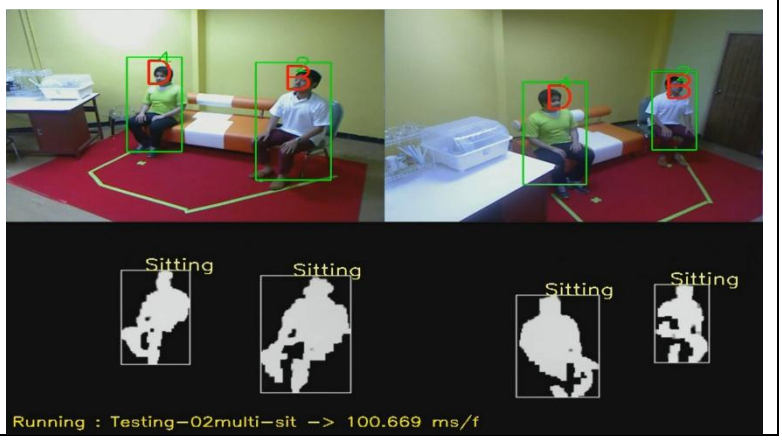
ตัวอย่างบุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก ในชุดทดสอบครั้งละสองคนชุดที่ 2 ได้แสดงไว้ในตารางที่ 4-24 และตัวอย่างการทำงานของระบบติดตามและรู้จำครั้งละสองคนในชุดทดสอบชุดที่ 2 ได้แสดงไว้ในตารางที่ 4-25

ตารางที่ 4-24 ตัวอย่างบุคคลที่เข้าทดสอบครั้งละสองคนและ Global ID สำหรับ Dataset # 2

ID	บุคคลที่เข้าทดสอบและGlobal ID ที่ถูก Assign ครั้งแรก	ID	บุคคลที่เข้าทดสอบและGlobal ID ที่ถูก Assign ครั้งแรก
#1 A		#3 C	
#2 B		#4 D	

ตารางที่ 4-25 ตัวอย่างการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละสองคนสำหรับ Dataset # 2

ตัวอย่างที่	ตัวอย่างเฟรมของการติดตามและจดจำ
1	
2	
3	

ตัวอย่างที่	ตัวอย่างเฟรมของการติดตามและจดจำ
4	

4.4.3 วิเคราะห์ผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง

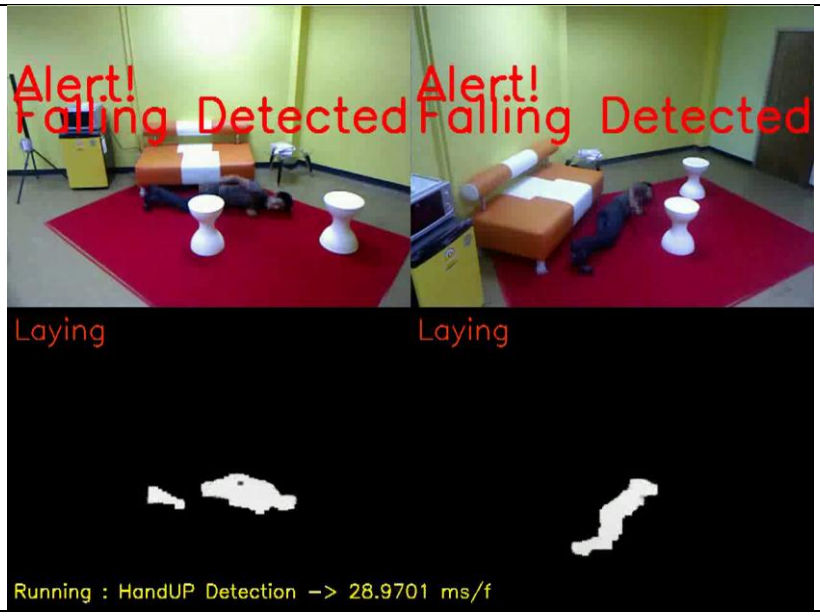
สำหรับการติดตามและจดจำตัวบุคคล จะใช้การวิเคราะห์ข้อมูลตำแหน่งและสี ซึ่งเป็นข้อมูลเบื้องต้นที่เด่นชัดที่สามารถนำมาใช้ในการแยกแยะแต่ละบุคคลเพื่อติดตามทั้งในกล้องเดียวกันและระหว่างกล้อง รวมไปถึงใช้ในการจดจำตัวบุคคลในเบื้องต้น ซึ่งการติดตามและจดจำตัวบุคคลมีความแม่นยำในกรณีที่เข้าไปในระบบครั้งละหนึ่งคนที่ 92.87% และกรณีที่เข้าไปในระบบครั้งละสองคนที่ 85.50% โดยจากผลการทดสอบจะมีการ Matching และการจดจำตัวบุคคล ที่ผิดพลาดบ้าง เนื่องจากปัจจัยของแสงที่ไม่เท่ากันในห้องและค่าสีที่ใกล้เคียงกันระหว่างบุคคล รวมไปถึงการเดินสลับกันเพื่อเปลี่ยนที่ ทำให้การติดตามสูญหายต้อง Matching ใหม่ แต่จากการทดสอบสามารถจับคู่บุคคลระหว่างกล้องได้มากกว่า 90% จึงทำให้ระบบการรู้จำท่าทางได้ข้อมูลที่ถูกต้องไป Fusion จากสองมุมมองและรู้จำท่าทางได้

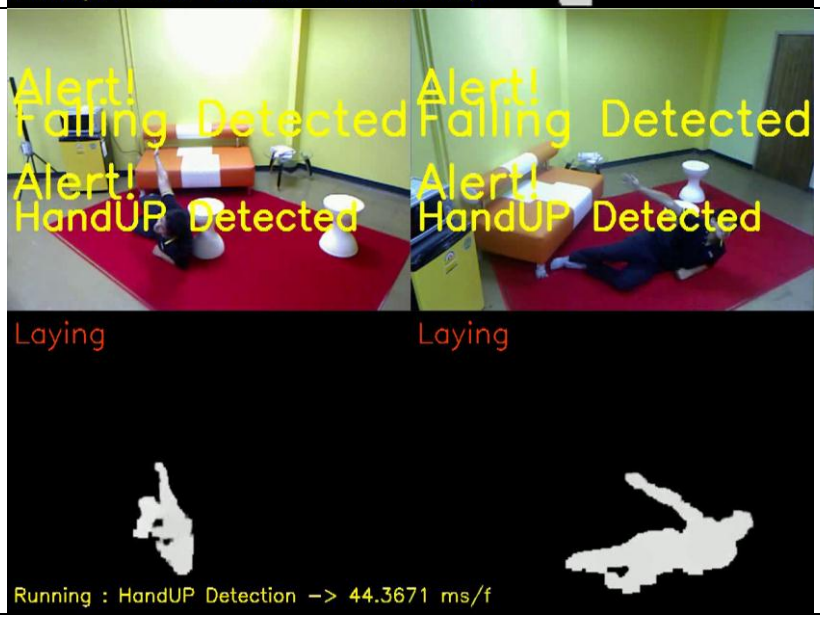
4.5 การทดสอบการตรวจจับท่าทางที่ผิดปกติ

4.5.1 กรณีศึกษาการล้ม

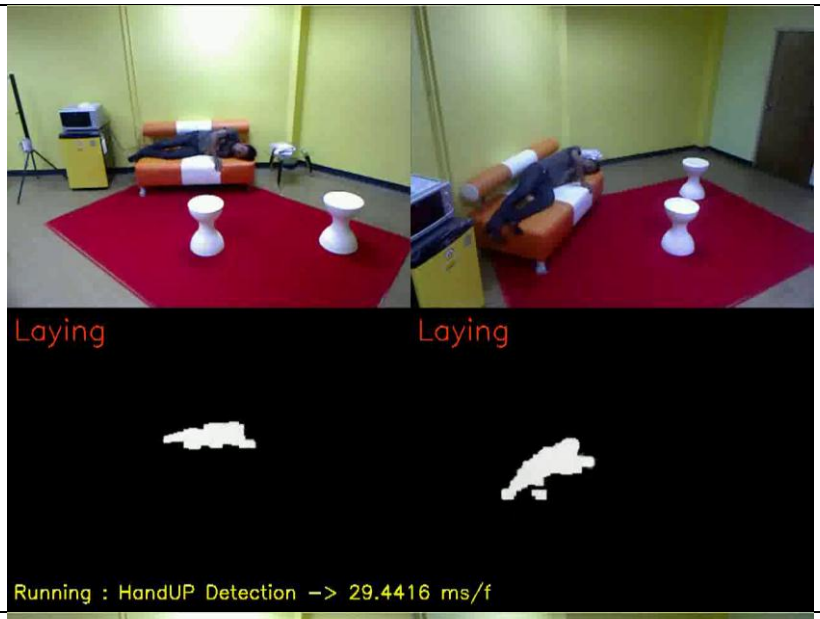
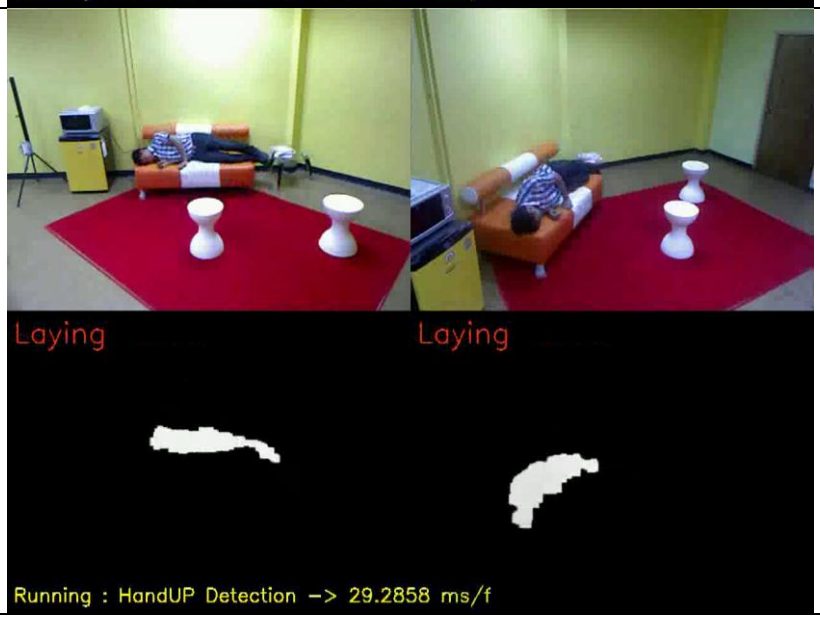
สำหรับการล้มจะมีปัจจัยเกี่ยวเนื่องกับการวิเคราะห์ท่าทางพื้นฐาน โดยการจับลำดับการเปลี่ยนแปลงจากท่าอื่น ๆเป็นท่านอน ซึ่งต้องใช้สถานที่มาเป็นตัวยืนยันว่าการเกิดนอนหรือการล้มโดยการสร้างจุดยกเว้นการตรวจจับขึ้นมา ประสิทธิภาพของการตรวจจับจึงขึ้นอยู่กับการรู้จำท่าทางที่เป็นการนอน ซึ่งสามารถรู้จำได้แม่นยำได้เท่ากับ 90.65% และท่าทางอื่นยังรายงานผลผิดพลาดเป็นท่านอนเพียงแค่ 0.05% ตามภาพประกอบที่ 4-12 ซึ่งในการตรวจจับการล้มจะมีฟังก์ชันการรอคอยให้มั่นใจว่าได้เกิดการล้มขึ้นจริง ๆ ซึ่งสามารถตั้งค่าระยะรอคอยให้มั่นใจว่าหลังจากเปลี่ยนลำดับท่าทางจากท่าทางอื่น ๆที่ไม่ใช่นอนมาเป็นนอนแล้วต้องนอน เท่ากับ N เฟรม จึงสามารถทำให้ระบบมีผลบวกปลอมที่น้อยลง ซึ่งตัวอย่างของการตรวจจับการล้มได้แสดงไว้ในตารางที่ 4-26 และ 4-27

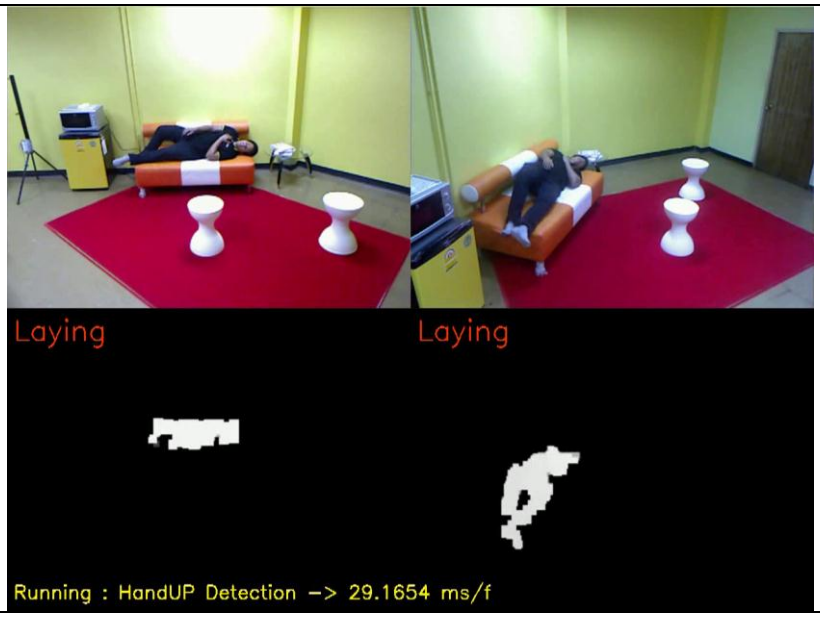
ตารางที่ 4-26 ตัวอย่างการตรวจจับการล้มหากเกิดการนอนอยู่นอกพื้นที่ที่ยกเว้น

ตัวอย่างที่	ตัวอย่างการตรวจจับการล้มหากเกิดการนอนอยู่นอกพื้นที่ที่ยกเว้น
1	

ตัวอย่างที่	ตัวอย่างการตรวจจับการล้มหากเกิดการนอนอยู่นอกพื้นที่ที่ยกเว้น
2	
3	

ตารางที่ 4-27 ตัวอย่างการไม่ถูกตรวจจับการล้มเมื่อล้มตัวลงในพื้นที่ที่ยกเว้น

ตัวอย่างที่	ตัวอย่างการไม่ถูกตรวจจับการล้มเมื่อล้มตัวลงในพื้นที่ที่ยกเว้น
1	
2	

ตัวอย่างที่	ตัวอย่างการไม่ถูกตรวจจับการล้มเมื่อล้มตัวลงในพื้นที่ที่ยกเว้น
3	

4.5.2 กรณีศึกษาการกระโดด

สำหรับการทดลองการตรวจจับการกระโดด จะทดลองกระโดดลงที่เดิม และกระโดดลงอีกที่ ทั้งกระโดดสูง และกระโดดต่ำ ๆ และมีทดสอบการกระโดดสองมุมมอง โดยจะใช้ผู้ทดสอบ 6 คน เป็นผู้ชาย 3 คน และผู้หญิง 3 คน ที่มีรูปร่างและความสูงต่างกัน โดยในการกระโดดแต่ละรูปแบบการทดสอบจะให้กระโดดกระโดดต่ำ ๆ 3 ครั้ง และกระโดดสูง 3 ครั้ง โดยมีรูปแบบการกระโดดดังนี้ กระโดดลงที่เดิมมุมมองที่ 1, กระโดดลงที่เดิมมุมมองที่ 2, กระโดดลงอีกที่มุมมองที่ 1, และกระโดดลงอีกที่มุมมองที่ 2 โดยแต่ละรูปแบบจะกระโดด 6 ครั้ง (กระโดดต่ำ ๆ 3 ครั้ง และกระโดดสูง 3 ครั้ง) ต่อ 1 คน รวมจำนวนกระโดดต่อ 1 คนคือ 24 ครั้ง จะได้จำนวนครั้งของการกระโดดทั้งหมด 144 ครั้ง และมีเหตุการณ์ที่ไม่ใช่การกระโดดเพื่อทดสอบความจำเพาะของระบบคือการยืนก่อนกระโดด 3 ครั้ง และรวมถึงการเปลี่ยนที่ก่อนการกระโดดอีกครั้ง 3 ครั้ง รวมเป็น 6 ครั้งในแต่ละรูปแบบการทดสอบ

การทดสอบจะใช้คอมพิวเตอร์ที่มีหน่วยประมวลผล CPU Intel Core i5 4590 ที่ความถี่ 3.30 GHz โดยใช้ OpenCV เป็น Library สำหรับประมวลผลด้าน Computer Vision และ CLNUI Library สำหรับติดต่อกับกล้อง Kinect และรับส่งข้อมูลภาพและข้อมูลความลึก โดยผลการทดสอบแสดงไว้ในตารางที่ 4-28 ซึ่งจะแสดงถึงค่า True Positive (TP), False Negative (FN), False Positive (FP) และ True Negative (TN)

ตารางที่ 4-28 ผลการทดสอบการตรวจจับการกระโดด

ผู้ทดสอบที่	TP/FN/ FP/TN								Summary of Object
	กระโดดลงที่เดิม				กระโดดลงอีกที่				
	มุมมองที่ 1		มุมมองที่ 2		มุมมองที่ 1		มุมมองที่ 2		
	Soft	Heavy	Soft	Heavy	Soft	Heavy	Soft	Heavy	TP/FN/ FP/TN
01	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	2/1/ 0/3	23/1/ 0/24
02	2/1/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 1/2	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	23/1/ 1/23
03	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	24/0/ 0/24
04	2/1/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	2/1/ 0/3	3/0/ 0/3	2/1/ 0/3	2/1/ 0/3	20/4/ 0/24
05	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	2/1/ 0/3	3/0/ 0/3	3/0/ 0/3	2/1/ 0/3	22/2/ 0/24
06	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	3/0/ 0/3	24/0/ 0/24
ผลรวม	16/2/ 0/18	18/0 /0/18	18/0/ 0/18	18/0/ 1/17	16/2/ 0/18	18/0/ 0/18	17/1 /0/18	15/3/ 0/18	136/8/ 1/143
ความไว	**88.89 %**	**100.0 %**	**100.0 %**	**100.0 %**	**88.89 %**	**100.0 %**	**94.44 %**	**83.33 %**	**94.44%**
ความจำเพาะ	**100.0 %**	**100.0 %**	**100.0 %**	**94.44 %**	**100.0 %**	**100.0 %**	**100.0 %**	**100.0 %**	**99.31%**

จากตารางที่ 4-28 ปรากฏว่าผลการทดลองการกระโดดมีค่าความไวเฉลี่ยเท่ากับ 94.44% ซึ่งค่านี้เป็นค่าที่บ่งบอกถึงโอกาสที่ระบบจะสามารถทำการตรวจจับการกระโดดได้ และค่าความจำเพาะเฉลี่ยเท่ากับ 99.31% ซึ่งค่านี้เป็นค่าที่บ่งบอกถึงโอกาสที่ระบบตอบผิดว่าเหตุการณ์อื่นที่ไม่ใช่การกระโดดเป็น

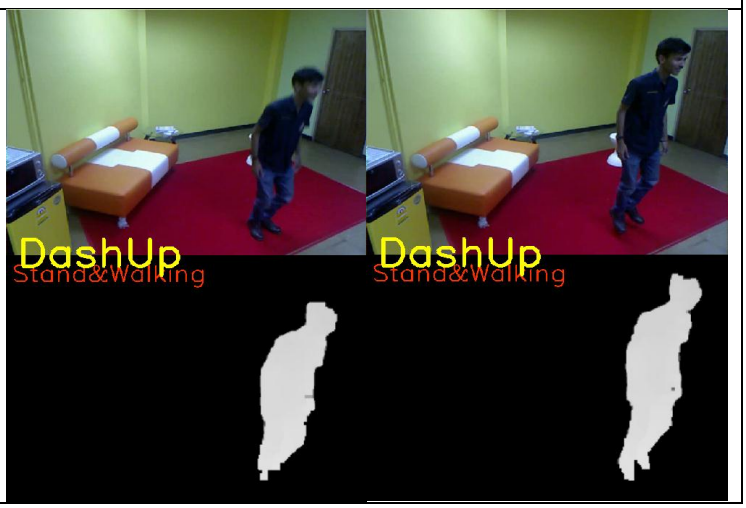
การกระโดด สำหรับค่า False Positive (FP) ที่เกิดขึ้นหนึ่งค่าในการทดสอบนั้นเป็นการเกิดขึ้นจากการกระโดดที่ต่อเนื่องกัน ที่อยู่ในระหว่างที่พักจะกระโดดครั้งใหม่ จะมีจังหวะดีดตัวขึ้นและจังหวะดิ่งลงที่เกิดขึ้นกับการกระโดดต่างครั้งกัน

จากผลการทดลองจะสรุปได้ว่าการกระโดดลงที่เดิมจะทำได้ดีกว่าการกระโดดลงอีกที่เนื่องจากการกระโดดอีกที่จะเป็นการกระโดดที่มีค่าการเปลี่ยนแปลงในแนวแกน y ต่ำมาก จึงทำให้ตรวจจับได้ยากกว่าหากกระโดดไม่สูงพอ และประการหนึ่งที่มีผลต่อระบบคือการบิดงอของมุมมองของกล้องในแต่ละมุมมีผลต่อการตรวจจับ เนื่องจากว่าในแต่ละมุมมองจะปรากฏระยะของการเคลื่อนที่ได้ไม่เท่ากันจึงทำให้มีผลต่อระบบในแบบจำลองคุณลักษณะการเคลื่อนไหวของวัตถุตาม Layer สำหรับตัวอย่างของการตรวจจับการกระโดดจะแสดงไว้ในตารางที่ 4-29

นอกเหนือจากนี้ยังมีการทดสอบค่าความจำเพาะกับชุดข้อมูลอื่น ๆที่เป็นการยกมือขึ้นและโบกมือ ซึ่งมีลักษณะการเคลื่อนไหวที่คล้าย ๆกันคือมีจังหวะขึ้นและลงของมือ ที่ใกล้เคียงคุณลักษณะการเคลื่อนไหวของการกระโดด แต่ไม่พบค่า False Positive ใด ๆออกมา

สำหรับการทดลองได้ บันทึกไว้ใน Youtube.com
ซึ่งสามารถเข้าถึงได้ที่
https://www.youtube.com/watch?v=CsbDWDKakxc

ตารางที่ 4-29 ตัวอย่างของการตรวจจับการกระโดด

ตัวอย่างที่	ภาพตัวอย่างของการตรวจจับการกระโดด
1	

ตัวอย่างที่	ภาพตัวอย่างของการตรวจจับการกระโดด
	
2	

ตัวอย่างที่	ภาพตัวอย่างของการตรวจจับการกระโดด
3	

4.5.3 กรณีศึกษาการโบกมือขอความช่วยเหลือ

โดยข้อมูลที่ใช้ในการทดสอบจะเป็นฉากของห้องนั่งเล่น ใช้คอมพิวเตอร์ที่มีหน่วยประมวลผล CPU Intel Core i5 4590 ที่ความถี่ 3.30 GHz โดยใช้ OpenCV เป็น Library สำหรับประมวลผลด้าน Computer Vision และ CLNUI Library สำหรับติดต่อกับกล้อง Kinect และรับส่งข้อมูลภาพและข้อมูลความลึก โดยได้แบ่งการประมวลผลในมุมมองเดี่ยวให้อยู่ใน Thread และประมวลผลพร้อม ๆกันแบบขนานโดยใช้ OpenMP Library ซึ่งระยะที่ใช้ในการทดสอบจะอยู่ในระยะ 3 ถึง 5.5 เมตร และจำกัดความสูงของข้อมูลจากผู้ทดสอบที่ 180 ซม. เพื่อให้ความลึกไม่มีการสูญเสียและได้ข้อมูลที่เป็นส่วนของร่างกายได้ครบถ้วนตั้งแต่ขาถึงศีรษะ โดยผลการทดสอบแสดงไว้ในตารางที่ 4-30 ซึ่งจะแสดงถึงค่า True Positive (TP), False Negative (FN), False Positive (FP)

ตารางที่ 4-30 ผลลัพธ์การทดสอบการโบกมือในท่าทางต่าง ๆ

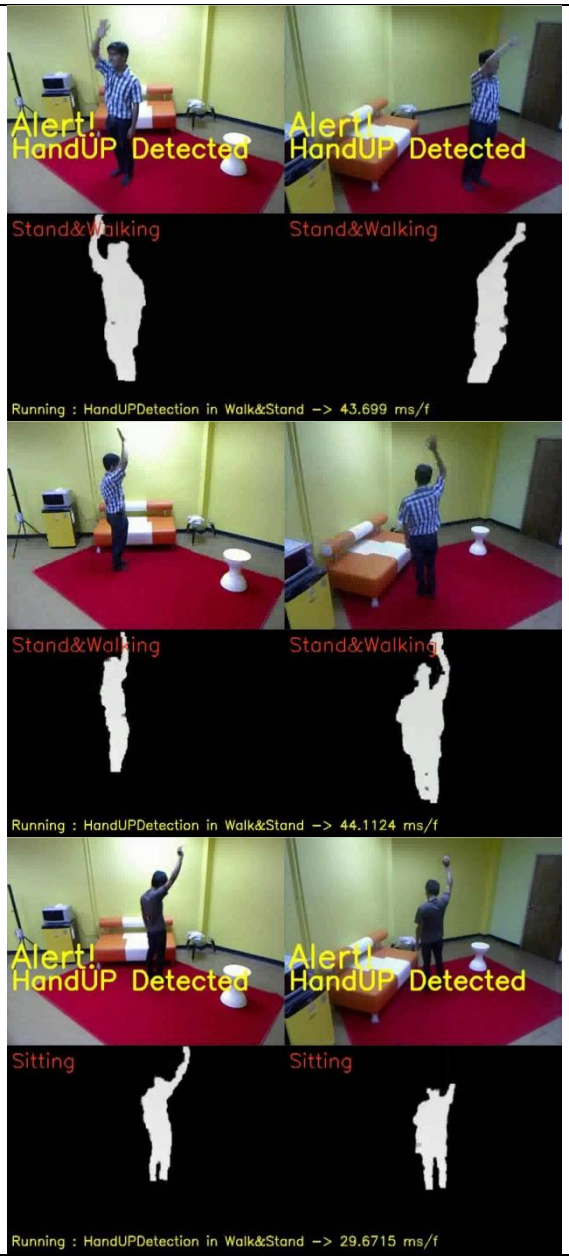
Action	TP/FN/FP				Summary of Action	
	ผู้ทดสอบ 01	ผู้ทดสอบ 02	ผู้ทดสอบ 03	ผู้ทดสอบ 04	TP/FN /FP	อัตราการตรวจจับ
ยืน / เดิน	7/1/0	8/0/0	8/0/0	8/0/0	31/1/0	96.88%
นั่ง	7/1/0	7/1/0	7/1/0	8/0/0	29/3/0	90.63%
ก้ม	8/0/0	7/1/0	8/0/0	8/0/0	31/1/0	96.88%
นอน	8/0/0	4/4/0	8/0/0	8/0/0	28/4/0	87.50%
ผลรวม	**30/2/0**	**28/4/0**	**31/1/0**	**32/2/0**	**119/9/0**	-
ค่าเฉลี่ย	**93.75%**	**87.50%**	**96.88%**	**100%**		**92.96%**

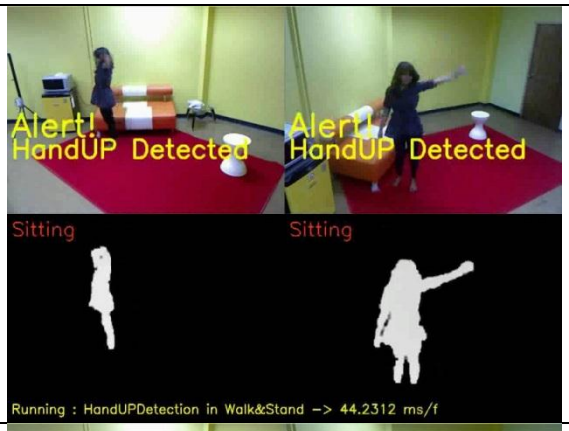
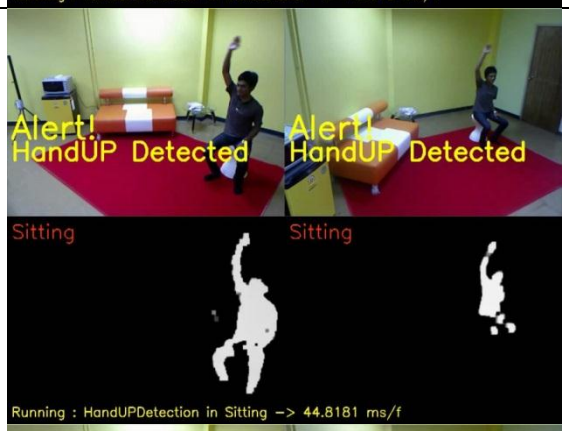
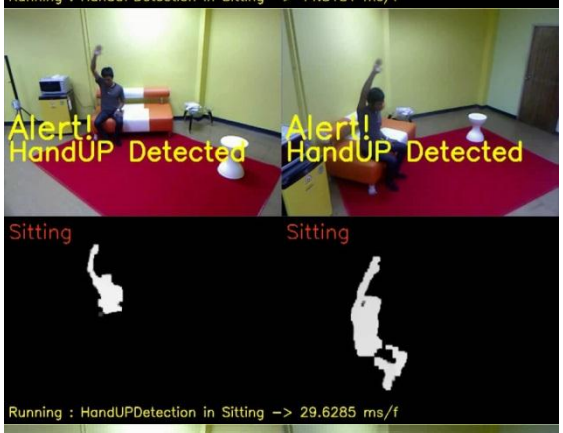
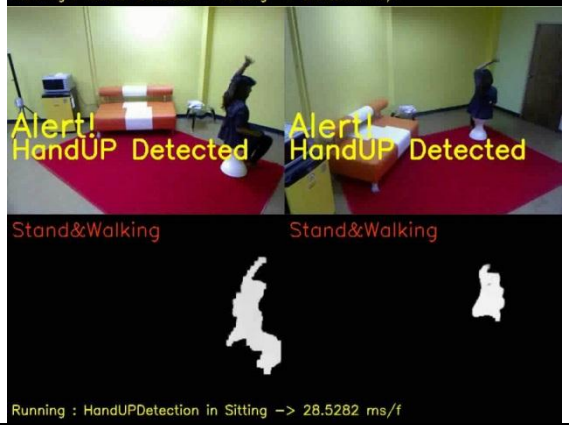
จากการทดลองมีอัตราการตรวจจับได้โดยเฉลี่ยทั้งหมด 92.96% และไม่การเตือนขึ้นมาหากไม่มีการโบกจริง (ไม่มีค่า False Positive) โดยการโบกมือในท่านอนทำได้น้อยสุด เนื่องจากปัจจัยของบุคคลที่ 2 ไม่ได้โบกมือโดยยกทั้งแขน และปัญหาจากความละเอียดของกล้องความลึกที่ไม่สามารถแยกระหว่างพนักพิงกับแขนของบุคคลได้ทำให้การโบกมือในท่านอนของบุคคลที่ 2 มีผลลัพธ์ที่น้อยกว่า ซึ่งขัดแย้งกับผลลัพธ์จากบุคคลอื่น ๆในท่านอน

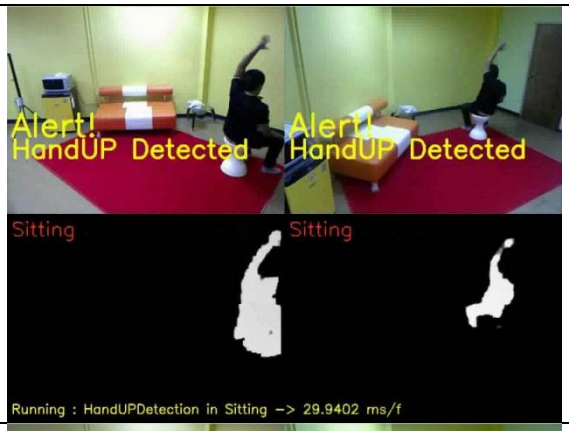
ซึ่งจากการวิเคราะห์เพื่อหาสาเหตุที่ทำให้เกิดค่า False Negative (FN) ก็ได้ทราบถึงสาเหตุหลัก ๆ 3 ประการ คือ (1) ข้อจำกัดในเรื่องมุมมอง ซึ่งในบางมุมมองไม่สามารถเห็นแขนที่ยกขึ้นมาโบกได้อย่างชัดเจนเนื่องจากมีการบดบัง, (2) ปัญหาจากเสถียรภาพของกล้องความลึกทำให้ในบางเฟรมไม่สามารถตรวจจับแขนได้, (3) ระยะเวลาและความเร็วในการโบกมือไม่เหมาะสม ซึ่งในการนำไปใช้จริงอาจจะต้องเลือกมุมมองในการตั้งกล้องที่มีค่าที่เหมาะสม และติดตั้งระบบตอบรับด้วยเสียง รวมถึงผู้ที่โบกมือขอความช่วยเหลือจะต้องพยายามโบกมือจนกว่าตอบรับ หากไม่ตอบรับให้เปลี่ยนมุมมองเล็กน้อยเพื่อช่วยระบบในการตรวจจับ และเป็นการช่วยให้ระบบสามารถตั้งค่าให้ทำงานได้โดยปราศจากการเตือนขึ้นมาโดยที่ไม่มีการโบกจริง (False Positive) โดยได้แสดงตัวอย่างการตรวจจับไว้ตามตารางที่ 4-31

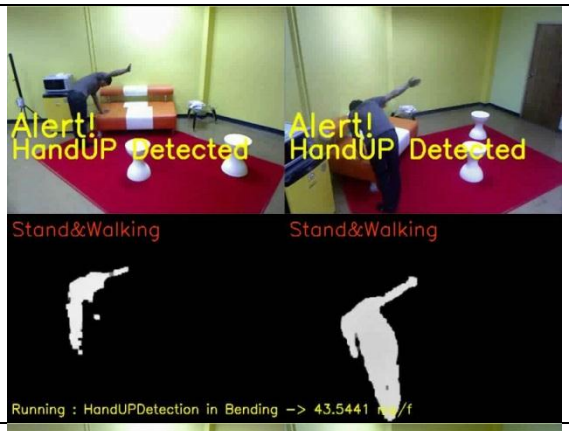
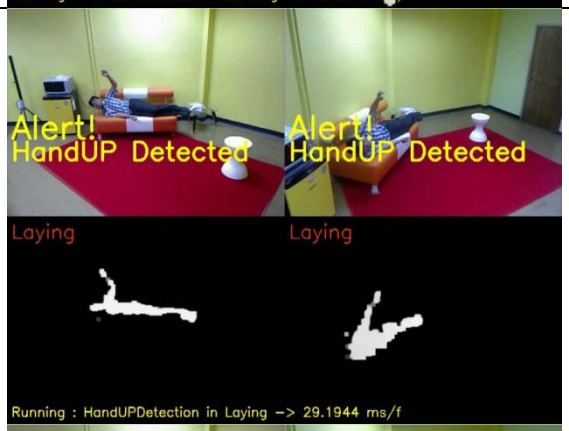
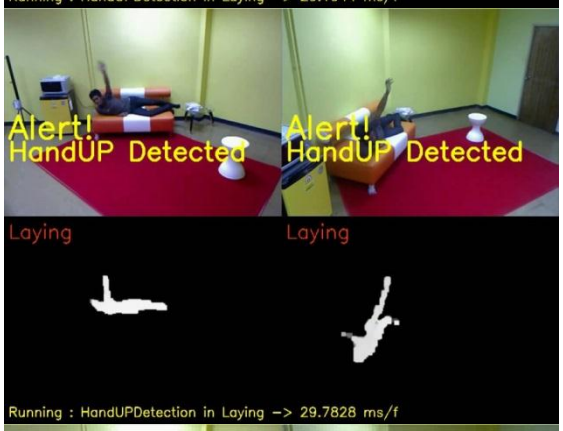
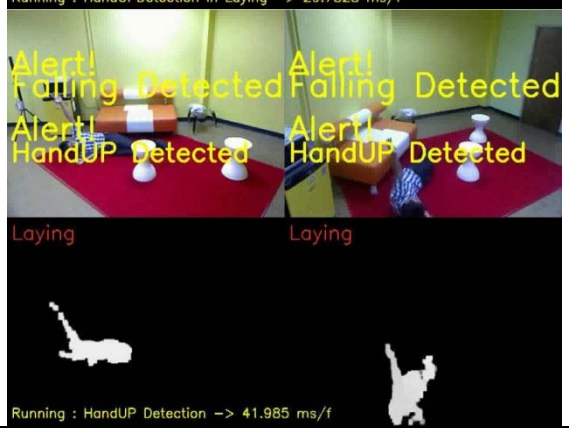
สำหรับการทดลองได้ บันทึกไว้ใน Youtube.com
ซึ่งสามารถเข้าถึงได้ที่
https://www.youtube.com/watch?v=X5x-98hszjc และ
https://www.youtube.com/watch?v=qVNLVV_8v8Y

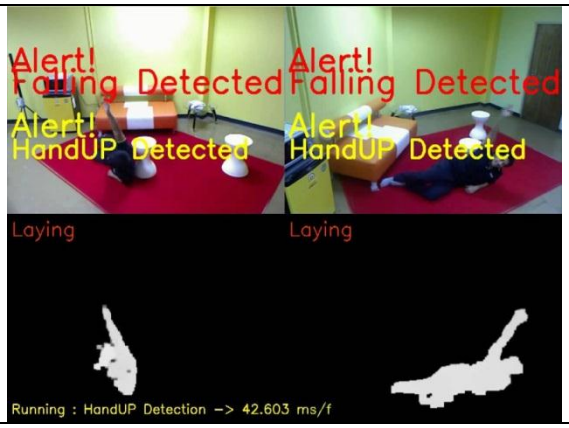
ตารางที่ 4-31 ตัวอย่างของการตรวจจับการโบกมือขอความช่วยเหลือ

ท่าทาง	ภาพตัวอย่างของการตรวจจับการโบกมือขอความช่วยเหลือ
ยืน / เดิน	

ท่าทาง	ภาพตัวอย่างของการตรวจจับการโบกมือขอความช่วยเหลือ
	
นั่ง	  

ท่าทาง	ภาพตัวอย่างของการตรวจจับการโบกมือขอความช่วยเหลือ
	
ก้ม	

ท่าทาง	ภาพตัวอย่างของการตรวจจับการโบกมือขอความช่วยเหลือ
	
นอน	  

ท่าทาง	ภาพตัวอย่างของการตรวจจับการโบกมือขอความช่วยเหลือ
	

4.5.4 วิเคราะห์ผลการทดสอบการตรวจจับท่าทางที่ผิดปกติ

สำหรับประสิทธิภาพของการตรวจจับการล้มนั้นขึ้นอยู่กับการรู้จำท่าทางที่เป็นการนอนที่มีความแม่นยำเท่ากับ 90.65% และท่าทางอื่นยังรายงานผลผิดพลาดเป็นท่านอนเพียงแค่ 0.05% อีกทั้งยังมีฟังก์ชันการรอคอยให้มั่นใจว่าได้เกิดการล้มขึ้นจริง ๆ จึงสามารถทำให้ระบบมีผลบวกปลอมที่น้อยลง ส่วนการกระโดดมีค่าความไวเฉลี่ยที่ 94.44% และค่าความจำเพาะเฉลี่ยที่ 99.31% ซึ่งการกระโดดลงที่เดิมจะทำได้ดีกว่าการกระโดดลงอีกที่เนื่องจากการกระโดดอีกที่จะเป็นการกระโดดที่มีค่าการเปลี่ยนแปลงในแนวแกน y ต่ำมาก จึงทำให้ตรวจจับได้ยากกว่าหากกระโดดไม่สูงพอ และประการหนึ่งที่มีผลต่อระบบคือมุมมองของกล้องที่บิดงอไปในแต่ละมุมมีผลต่อการตรวจจับ เนื่องจากว่าในแต่ละมุมมองจะปรากฏระยะของการเคลื่อนที่ได้ไม่เท่ากันจึงทำให้มีผลต่อระบบในแบบจำลองคุณลักษณะการเคลื่อนไหวของวัตถุตาม Layer ส่วนการตรวจจับการโบกมือมีอัตราการตรวจจับได้โดยเฉลี่ยทั้งหมด 92.96% และไม่การเตือนขึ้นมาหากไม่มีการโบกจริง (ไม่มีค่า False Positive) โดยการโบกมือในท่านอนทำได้น้อยสุด และปัญหาจากความละเอียดของกล้องความลึกที่ไม่สามารถแยกระหว่างพนักพิงกับแขนของบุคคลได้ ส่วนสาเหตุที่ทำให้เกิดค่า False Negative ก็ได้ทราบถึงสาเหตุหลัก ๆ 3 ประการ คือ (1) ข้อจำกัดในเรื่องมุมมองซึ่งในบางมุมมองไม่สามารถเห็นแขนที่ยกขึ้นมาโบกได้อย่างชัดเจนเนื่องจากมีการบดบัง, (2) ปัญหาจากเสถียรภาพของกล้องความลึกทำให้ในบางเฟรมไม่สามารถตรวจจับแขนได้, (3) ระยะเวลาและความเร็วในการโบกมือไม่เหมาะสม ซึ่งในการนำไปใช้จริงอาจจะต้องเลือกมุมมองในการตั้งกล้องที่มีค่าที่เหมาะสม และติดตั้งระบบตอบรับด้วยเสียง รวมถึงผู้ที่โบกมือขอความช่วยเหลือจะต้องพยายามโบกมือจนกว่าตอบรับ หากไม่ตอบรับให้เปลี่ยนมุมมองเล็กน้อยเพื่อช่วย

ระบบในการตรวจจับ และเป็นการช่วยให้ระบบสามารถตั้งค่าให้ทำงานได้โดยปราศจากการเตือนขึ้นมาโดยที่ไม่มีการโบกจริง

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

ในบทนี้จะกล่าวถึงบทสรุปของการวิจัยและการพัฒนาวิธีติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง ที่ผนวกเข้ากับระบบการรู้จำท่าทางจากหลายมุมมองประกอบไปด้วยการฟิวชันข้อมูลในระดับสูงจากหลายมุมมอง และการฟิวชันฟีเจอร์ในระดับล่างจากหลายมุมมอง ซึ่งจะใช้ข้อมูลจากการจับคู่บุคคลที่ตรงกันจากการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมองเพื่อนำฟีเจอร์ของบุคคลเดียวกันมาใช้ในการรู้จำท่าทางต่อไป หลังจากนั้นคำตอบที่เป็นท่าทางของบุคคลจะมีส่วนในการติดสินใจของระบบตรวจจับการล้ม และระบบตรวจจับการกระโดด นอกจากนี้ยังมีระบบตรวจจับการโบกมือขอความช่วยเหลือ ที่เป็นส่วนที่อิสระออกมาจากส่วนอื่น ซึ่งทั้งหมดนี้จะใช้ข้อมูลภาพสีและความลึกจากหลายมุมมอง

5.1 สรุปผลการวิจัยและอภิปรายผล

สำหรับงานวิจัยนี้ได้วิจัยและพัฒนาวิธีการวิเคราะห์ท่าทางพื้นฐานได้แก่ การยืน / การเดิน การนั่ง การก้ม และการนอน โดยใช้ข้อมูลภาพสีและความลึกจากหลายมุมมอง ซึ่งผู้วิจัยได้นำเสนอการฟิวชันข้อมูลในระดับสูงจากหลายมุมมอง โดยมีแนวความคิดของการฟิวชันข้อมูลในระดับคำตอบได้มาจากการที่สังเกตการณ์รู้จำท่าทางในหลากหลายมุมมองที่ส่องไปยังมนุษย์ จะพบว่าประสิทธิภาพในการรู้จำต่าง ๆขึ้นอยู่กับมุมมอง โดยงานวิจัยนี้จะเพิ่มความถูกต้องของการรู้จำโดยสร้างฟังก์ชันวัดค่าความน่าเชื่อถือของคำตอบ โดยใช้ผลลัพธ์ของการรู้จำมุมมองและผลจากการทดลองทดลองเชิงประจักษ์เป็นเกณฑ์ ซึ่งจากผลการทดสอบการรู้จำท่าทางโดยการฟิวชันปรากฏว่ามีความแม่นยำเฉลี่ยเพิ่มมากขึ้นจากการรู้จำท่าทางมุมมองเดี่ยวถึง 11.86% และ 16.66% ของฟิวชันแบบจำลองพื้นฐานและซับซ้อนตามลำดับ และสามารถเพิ่มความแม่นยำในหลาย ๆ มุมมองและท่าทางโดยเฉพาะท่านอนด้านหน้าที่สามารถเพิ่มความแม่นยำได้ถึง 98.17% นอกจากการฟิวชันข้อมูลในระดับสูงแล้วผู้วิจัยยังได้พัฒนาวิธีการฟิวชันฟีเจอร์ในระดับล่างจากหลายมุมมอง โดยใช้แบบจำลองการฟิวชันแบบเลเยอร์ ที่แบ่งพื้นที่เพื่อ Encode ฟีเจอร์ในเลเยอร์ในมุมมองเดี่ยวและนำไปฟิวชันเพื่อปรับปรุงเป็นฟีเจอร์ใหม่โดยใช้ข้อมูลจากหลายมุมมอง จากนั้นเวกเตอร์ของฟีเจอร์เหล่านี้จะถูกนำไปใช้ในขั้นตอนการเรียนรู้จดจำและแยกแยะ เพื่อสามารถจะนำมาใช้ในการรู้จำท่าทางต่อไป ผู้วิจัยได้ทำการทดสอบเปรียบเทียบประสิทธิภาพของการรู้จำท่าทางจากมุมมองเดี่ยวและหลายมุมมองจากชุดข้อมูล PSU และ i3DPost ซึ่งชุดข้อมูล PSU มีค่าเฉลี่ยของความแม่นยำในมุมมองเดี่ยวเท่ากับ 92.50% และ 90.63% ขณะที่สองมุมมองมี

ค่าเฉลี่ยของความแม่นยำอยู่ที่ 95.32% ขณะที่ค่าเฉลี่ยของความแม่นยำในชุดข้อมูล i3DPost เป็น 89.03%, 93.00%, 91.33%, 92.30%, 92.56% (เรียงจากจำนวนของมุมมองที่ 1-6) ซึ่งก็มีแนวโน้มที่เพิ่มขึ้นจากมุมมองเดี่ยว จึงสรุปได้ว่าจำนวนของมุมมองเพียงแค่ 2 หรือ 3 จากมุมมองที่แตกต่างกันก็เพียงพอแล้วสำหรับการรู้จำท่าทางในแบบจำลองของผู้วิจัย โดยที่การมีมุมมองที่มากขึ้นอาจจะมีข้อมูลที่มากเกินจำเป็น และยังเป็นการสะสมสิ่งรบกวนได้ นอกจากนี้ผู้วิจัยได้ทำการทดสอบความทนทานของแบบจำลอง อันได้แก่ (1) ทดสอบกับชุดข้อมูล PSU ในฉากห้องนั่งเล่น ซึ่งสอนข้อมูลฉากห้องทำงาน มีความแม่นยำที่สูงที่สุดจากค่าเฉลี่ย 95.32% (2) ทดสอบกับชุดข้อมูล PSU ซึ่งสอนข้อมูลโดย NW-UCLA มีความแม่นยำที่สูงที่สุดจากค่าเฉลี่ยที่ 93.44% ซึ่งได้ผลลัพธ์ค่อนข้างดีในระดับหนึ่ง, (3) ทดสอบกับชุดข้อมูล NW-UCLA ซึ่งสอนข้อมูลโดย PSU มีความแม่นยำโดยเฉลี่ยอยู่เพียงแค่ 86.40% โดยที่ชุดข้อมูล NW-UCLA มีการวางมุมของกล้องและช่วงระยะจากกล้องถึงตัวบุคคลมีความแปรปรวนมาก รวมไปถึงมีความแตกต่างกันชัดเจนกับชุดข้อมูล PSU, (4) ทดสอบกับชุดข้อมูล i3DPost ซึ่งสอนข้อมูลโดย PSU มีความแม่นยำโดยเฉลี่ยที่ 72.03% เนื่องจากท่านั่งของชุดข้อมูล i3DPost เป็นท่านั่งยองในอากาศจึงทำให้ท่านั่งถูกทำนายเป็นท่ายืน ในทางที่กลับกัน ท่ายืน/เดิน และท่าก้มให้ผลลัพธ์ที่ค่อนข้างดีถึง 96.40% และ100% ซึ่งจากผลการทดสอบที่ใช้ข้อมูลที่สอนและทดสอบต่างชุดข้อมูลกัน ผู้วิจัยให้ความเห็นว่าแบบจำลองในการรู้จำท่าทางนี้มีความทนทานต่อการเปลี่ยนแปลงในระดับหนึ่ง แต่มุมกล้องที่ติดตั้งต่างกันจากข้อมูลมาก ๆ รวมไปถึงท่าทางที่นอกเหนือจากการสอนข้อมูล ก็มีผลต่อประสิทธิภาพของระบบบ้าง นอกจากนี้ผู้วิจัยยังได้ทดสอบการทำมุมกันของกล้องที่องศาแตกต่างกันจากการสอนข้อมูลที่กล้องตั้งฉากกัน ซึ่งผลปรากฏว่าระบบก็ยังสามารถรู้จำได้ดีอยู่ แต่ความแม่นยำในบางท่าทางจะลดลงบ้าง โดยภาพรวมจะลดลงเพียงเล็กน้อย ซึ่งจากการเปรียบเทียบการรู้จำท่าทางจากหลายมุมมอง เห็นได้ว่างานวิจัยวิธีการฟิวชันฟีเจอร์ในระดับล่างจากหลายมุมมองค่อนข้างมีประสิทธิภาพ เนื่องจากว่าแบบจำลองมีความซับซ้อนที่น้อย เน้นความเป็นส่วนตัวในการสังเกตการณ์ที่สูง เนื่องจากใช้ภาพความลึกและคุณสมบัติอื่น ๆ เช่น ความยืดหยุ่นสามารถเพิ่มขยายระบบได้ และความทนทานซึ่งมีระดับที่ค่อนข้างดี หรือไม่น้อยกว่าและไม่ต้องการการสอบเทียบค่าของกล้องในการติดตั้ง

นอกจากนี้งานวิจัยยังนำเสนอการตรวจจับท่าทางที่ผิดปกติจากหลายมุมมอง โดยเป็นการนำแบบจำลองและคำตอบจากการรู้จำท่าทางพื้นฐาน ได้แก่ ยืน / เดิน นั่ง ก้ม นอน มาประยุกต์ต่อยอดในการใช้ตรวจจับเหตุการณ์ที่ผิดปกติโดยที่ใช้แค่ Rule-based ขั้นพื้นฐานเท่านั้น โดยจะทำการทดสอบในฐานข้อมูล PSU แค่เบื้องต้นเท่านั้น สำหรับการตรวจจับการกระโดดนั้น ผู้วิจัย

ได้นำแบบจำลองการฟิวชันแบบเลเยอร์มาติดตามหาการเคลื่อนที่โดยรวมของบุคคลเพื่อนำไปตรวจจับจังหวะดีดตัวขึ้นแล้วดิ่งลง ที่เป็นจังหวะที่สำคัญของการกระโดด โดยที่นำคำตอบจากการรู้จำท่าทางพื้นฐานมาใช้ในการตัดค่าผลบวกปลอมจากการตรวจจับ เนื่องจากท่ากระโดดต้องมีพื้นฐานมาจากท่ายืนเสมอ สำหรับการทดสอบการตรวจจับการกระโดด จะทดสอบกระโดดลงที่เดิม และกระโดดลงอีกที่ ทั้งกระโดดสูง และกระโดดต่ำ ๆ และมีทดสอบการกระโดดสองมุมมอง โดยจะใช้ผู้ทดสอบ 6 คน เป็นผู้ชาย 3 คน และผู้หญิง 3 คน ที่มีรูปร่างและความสูงต่างกัน โดยในการกระโดดแต่ละรูปแบบการทดสอบจะให้กระโดดต่ำ ๆ 3 ครั้ง และสูง 3 ครั้ง ซึ่งจะได้จำนวนครั้งของการกระโดดทั้งหมด 144 ครั้ง และมีเหตุการณ์ที่ไม่ใช่การกระโดดเพื่อทดสอบความจำเพาะของระบบคือการยืนก่อนกระโดด 3 ครั้ง และรวมถึงการเปลี่ยนที่ก่อนการกระโดดอีกครั้ง 3 ครั้ง รวมเป็น 6 ครั้งในแต่ละรูปแบบการทดสอบ จากการทดสอบการกระโดดจะมีค่าความไวของการตรวจจับเฉลี่ยที่ 94.44% และค่าความจำเพาะเฉลี่ยที่ 99.31% และสำหรับค่า False Positive (FP) ที่เกิดขึ้นหนึ่งค่าจากทั้งหมดในการทดสอบซึ่งจากผลการทดสอบสรุปได้ว่าการกระโดดลงที่เดิมจะทำได้ดีกว่าการกระโดดลงอีกที่เนื่องจากการกระโดดอีกที่จะเป็นการกระโดดที่มีค่าการเปลี่ยนแปลงในแนวแกน y ต่ำมาก จึงทำให้ตรวจจับได้ยากกว่าหากกระโดดไม่สูงพอ และประการหนึ่งที่มีผลต่อระบบคือมุมมองของกล้องในแต่ละมุมก็มีผลต่อการตรวจจับเนื่องจากว่าในแต่ละมุมมองจะปรากฏระยะของการเคลื่อนที่ได้ไม่เท่ากันจึงทำให้มีผลต่อระบบในแบบจำลองคุณลักษณะการเคลื่อนไหวของวัตถุตาม Layer สำหรับตัวอย่างของการตรวจจับการกระโดด ส่วนการตรวจจับการล้มจะใช้การประยุกต์ต่อยอดจากการวิเคราะห์ท่าทางพื้นฐาน โดยการจับลำดับการเปลี่ยนแปลงจากท่าอื่น ๆเป็นท่านอน และใช้สถานที่มาเป็นตัวยืนยันว่าการเกิดนอนหรือการล้ม โดยในงานวิจัยนี้ใช้การสร้างจุดยกเว้นการตรวจจับ ซึ่งเป็นบริเวณที่บุคคลจะไปนอนได้ อีกทั้งยังมีฟังก์ชันการรอคอยให้มั่นใจว่าได้เกิดการล้มขึ้นจริง ๆ จึงสามารถทำให้ระบบมีผลบวกปลอมที่น้อยลง สำหรับประสิทธิภาพของการตรวจจับการล้มนั้นขึ้นอยู่กับการรู้จำท่าทางที่เป็นการนอนที่มีความแม่นยำเท่ากับ 90.65% และท่าทางอื่นยังรายงานผลผิดพลาดเป็นท่านอนเพียงแค่ 0.05% ส่วนการตรวจจับการโบกมือจะประยุกต์ใช้เทคนิคการวิเคราะห์ภาพฉายไบนารี่ในแต่ละมุมมองและใช้การตอบสนองโดยเสียง เพื่อให้ผู้ใช้รับรู้ได้ว่าระบบได้ตรวจจับแล้วหรือไม่ ซึ่งจะสามารถจะปรับแต่งทำให้ระบบมีค่าผลบวกปลอมที่น้อยลง โดยมีอัตราการตรวจจับการโบกมือได้โดยเฉลี่ยทั้งหมด 92.96% และไม่การเตือนขึ้นมาหากไม่มีการโบกจริง (ไม่มีค่า False Positive)

โดยสรุป หัวข้อต่าง ๆที่ผู้วิจัยพัฒนาขึ้นมาไม่ว่าจะเป็นการรู้จำท่าทางพื้นฐานของมนุษย์ ประกอบด้วย ยืน / เดิน นั่ง นอน และก้ม สามารถนำไปสู่การรู้จำการเข้าใจท่าทางที่ซับซ้อนกว่า กิจกรรม และพฤติกรรมของมนุษย์ ซึ่งในงานวิจัยนี้ได้ประยุกต์บางส่วนไปใช้ในตรวจจับการล้ม และการกระโดด และงานวิจัยนี้ยังได้นำเสนอวิธีการติดตามและจดจำตัวบุคคลจากหลายมุมมอง ที่สามารถติดตามและจดจำตัวบุคคลที่เข้าออกในสถานที่หนึ่ง ๆได้ โดยสิ่งเหล่านี้สามารถนำไป ประยุกต์ใช้กับระบบรักษาความปลอดภัย รวมถึงการดูแลด้านสุขภาพของบุคคล โดยเฉพาะอย่าง ยิ่งในการเฝ้าระวังผู้สูงอายุที่อยู่บ้านคนเดียว ซึ่งเหมาะกับสถานการณ์ในปัจจุบันที่กำลังเข้าสู่สังคม ผู้สูงอายุ อีกทั้งระบบที่ผู้วิจัยพัฒนาขึ้นยังใช้ข้อมูลภาพความลึกจากหลายมุมมองเป็นหลักซึ่งจะมี ความทนทานต่อสภาวะแสงที่เปลี่ยนแปลงไปรวมไปถึงการเปลี่ยนแปลงของมุมมอง ซึ่งจะทำให้ได้ ข้อมูลหลายรูปแบบจากหลายมุมมองที่มากขึ้น โดยจะทำให้มีประสิทธิภาพในการวิเคราะห์และ รู้จำมากยิ่งขึ้น

5.2 ข้อเสนอแนะ

สำหรับในเรื่องของการรู้จำท่าทางโดยการฟิวชันข้อมูลในระดับสูงจากหลายมุมมอง ผู้วิจัย เห็นว่าจะสามารถปรับปรุงให้ผลลัพธ์ของการฟิวชันดีขึ้นได้โดยการเพิ่มประสิทธิภาพของตัวรู้จำ มุมมอง ซึ่งปัจจุบันทำได้เพียง 84.90% ส่วนการรู้จำท่าทางโดยการฟิวชันฟีเจอร์ในระดับล่างจาก หลายมุมมอง สามารถนำไปพัฒนาและทดสอบหาฟีเจอร์อื่น ๆซึ่งอาจจะให้ผลลัพธ์ที่ดีขึ้น และใช้ ฟีเจอร์ที่เป็นเชิงพื้นที่และเวลา ซึ่งอาจจะนำไปสู่รู้จำท่าทางที่ซับซ้อนได้ดีมากขึ้น รวมไปถึงการ ทดสอบการรู้จำโดยใช้วิธีการจำแนกประเภทของข้อมูลอื่น ๆ และการตัดสินใจคำตอบเป็นท่าทาง ควรจะมี Unknown Class เพื่อจัดการท่าทางที่ไม่อยู่ในขอบเขตในการรู้จำในงานวิจัยนี้ นอกจากนี้ การนำชุดข้อมูลที่มีระยะการทำงานที่เป็นภายนอกอาคารมาทดสอบ (ไม่อยู่ในขอบเขตของ งานวิจัยนี้ที่เป็นในอาคาร) ซึ่งหากระยะจากกล้องเปลี่ยนไปและทำให้วัตถุเล็กลง จะทำให้สามารถ ประเมินความสามารถของแบบจำลองที่ได้นำเสนอได้ดียิ่งขึ้น

สำหรับการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมองที่ใช้การวิเคราะห์ข้อมูล ตำแหน่งและสี ซึ่งเป็นข้อมูลเบื้องต้นที่เด่นชัดที่สามารถนำมาใช้ในการแยกแยะแต่ละบุคคลได้ แต่ ก็ยังไม่สามารถทำได้ดีเท่าที่ควร เนื่องจากปัจจัยของแสงที่ไม่เท่ากันในห้องและค่าสีที่ใกล้เคียงกัน ระหว่างบุคคล ซึ่งอาจจะต้องพัฒนาแบบจำลองของสีใหม่ที่ตัดสิ่งรบกวนเรื่องแสงที่เปลี่ยนไป รวมถึงการนำรูปร่าง พื้นผิว หรืออื่น ๆที่เป็นองค์ประกอบที่สามารถรู้จำบุคคลได้ดีกว่าสีมาใช้ใน การติดตามและแยกแยะตัวบุคคล รวมไปถึงการเดินสลับกันเพื่อเปลี่ยนที่ ทำให้การติดตามสูญ หายต้องจับคู่ใหม่ ที่น่าจะสามารถแก้ปัญหาด้วยการใส่แบบจำลองการทำนายบุคคลเมื่อวัตถุ หายไปจากการเคลื่อนที่เข้าไป

นอกจากนี้เซ็นเซอร์ความลึกที่ใช้ในงานวิจัยนี้เป็น Kinect version 1 ซึ่งมีความแปรปรวนและการสูญหายของค่าความลึกมาก ซึ่งหากเปลี่ยนเป็นเซ็นเซอร์ภาพความลึกอื่น ๆที่เสถียรกว่านี้อาจจะทำให้ระบบมีประสิทธิภาพที่ดีขึ้น เช่น Kinect version 2, Orbbec Depth Sensor, Intel RealSense เป็นต้น

เอกสารอ้างอิง

[1] N. Noorit and N. Suvonvorn, "Human Activity Recognition System Based-on Sequential Logic Circuits and Statistical Models", *GSTF Journal on Computing (JoC)* Vol.5 No.3, May 2017.

[2] N. Noorit and N. Suvonvorn, (2015), "Human Activity Recognition from Basic Actions Using Graph Similarity Measurement", *In Proceeding of 12th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Hatyai, 22-24 July, pp.7-11.

[3] S. Ruangniam and N. Suvonvorn, (2017), "Rule-based Activity Classification Technique using Actions and Objects Information", *In Proceeding of International Conference on Electronics, Information, and Communication (ICEIC)*, Phuket, 11-14 January, pp.1-4.

[4] P. Chawalitsittikul and N. Suvonvorn, "Profile-based human action recognition using depth information," *in Proceeding of Advances Computer Science and Engineering ACTA Press*, pp. 376-380, 2012.

[5] Z. He and X. Bai, "A wearable wireless body area network for human activity recognition," *in Proceeding of IEEE Ubiquitous and Future Network*, pp. 115-119, 2014.

[6] C. Zhu and W. Sheng, "Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living," *IEEE transactions on systems, man, and cybernetics—part a: systems and humans*, vol. 41, no. 3, pp.569-573, 2011.

[7] J. S. Sheu, G. S. Huang, W. C. Jheng et al., "Design and implementation of a three dimensional pedometer accumulating walking or jogging motions" *in Proceeding of IEEE International symposium on computer, computer and control*, pp. 828-831, 2014.

[8] R. S. Zonouz, H. M. Tehran, and R. Rahmani, "Smartphone-centric human posture monitoring system", *in Proceeding of IEEE Canada International Humanitarian Technology Conference*, pp. 1-4, 2014.

[9] X. Yin, W. Shen, J. Samarabandu et al., "Human activity detection based on multiple smart phone sensors and machine learning algorithms", *in Proceeding of IEEE 19th International Conference on Computer Supported Cooperative Work in Design*, pp. 582-587, 2015.

[10] C. A. Siebra, B. A. Sá, T. B. Gouveia et al., "A neural network based application for remote monitoring of human behaviour", *in Proceeding of IEEE Computer Vision and Image Analysis Applications*, pp. 1-6, 2015.

[11] C. Pham, "MobiRAR: real-time human activity recognition using mobile devices," *in Proceeding of IEEE Seventh International Conference on Knowledge and Systems engineering*, pp. 144-149, 2015.

[12] G. M. Weiss, J. L. Timko, C. M. Gallagher et al., "Smartwatch-based activity recognition: a machine learning approach," *in Proceeding of IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 426-429, 2016.

[13] W. Ye and B. Xiang-yu, "Research of fall detection and alarm applications for the elderly," *in Proceeding of IEEE International Conference on Mechatronic Sciences, Electric Engineering and Computer*, pp. 615-619, 2013.

[14] Y. Ge and B. Xu, "Detecting falls using accelerometers by adaptive thresholds in mobile devices," *Journal of computers in academy publisher*, vol. 9, no. 7, pp. 1553-1559, 2014.

[15] W. Liu, Y Lou, J Yan et al., "Falling monitoring system based on multi axial accelerometer," *in Proceeding of IEEE the 11th World Congress on Intelligent Control and Automation*, pp. 7-12, 2014.

[16] J. Yin, Q. Yang, and J. J. Pan, "Sensor-based abnormal human-activity detection," *IEEE transactions on knowledge and data engineering*, vol. 20, no. 8, 2008.

[17] B. X.Nie, C. Xiong, and S. C. Zhu, "Joint action recognition and pose estimation from video," *in Proceeding of IEEE Computer Vision and Pattern Recognition*, pp. 1293-1301, 2015.

[18] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal quads: human action recognition using joint quadruples," *in Proceeding of IEEE 22nd International Conference on Pattern Recognition*, pp. 4513-4518, 2014.

[19] E. Ohn-Bar and M. M. Trivedi, "Joint angles similiarities and HOG2 for action recognition," *in Proceeding of IEEE Computer Vision and Pattern Recognition Workshops*, pp. 465-470, 2013.

[20] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," *in Proceeding of IEEE Computer Vision and Pattern Recognition*, pp. 588-595, 2014.

[21] V. Parameswaran, and R. Chellappa, "View invariance for human action recognition," *International Journal of Computer Vision (IJCV)*, vol. 66, no. 1, pp. 83-101, 2006.

[22] G. Lu, Y. Zhou, X. Li et al., "Efficient action recognition via local position offset of 3D skeletal body joints," *Journal of Multimedia Tools and Applications*, vol. 75, pp. 3479-3494, 2015.

[23] A. Tejero-de-Pablos, Y. Nakashima, N. Yokoya et al., "Flexible human action recognition in depth video sequences using masked joint trajectories," *EURASIP Journal on Image and Video Processing*, vol. 20, pp. 1-12, 2016.

[24] E. Cippitelli, S. Gasparrini, Ennio Gambi et al., "A human activity recognition system using skeleton data from RGBD sensors," *Journal of Computational Intelligence and Neuroscience, Hindawi*, vol. 2016, 2016.

[25] P. Peursum, H. Bui, S. Venkatesh et al., "Robust Recognition and Segmentation of Human Actions Using HMMs with Missing Observations," *EURASIP Journal on Applied Signal Processing, Hindawi*, vol. 2005, no. 13, pp. 2110-2126, 2005.

[26] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Journal of Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249-257, 2006.

[27] H. Wang, A. Kläser, C. Schmid et al., "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision (IJCV)*, vol. 103, pp. 60-79, 2013.

[28] P. Matikainen, M. Hebert, and R. Sukthankar, "Trajectons: Action recognition through the motion analysis of tracked features," *in Proceeding of IEEE 12th International Conference on Computer Vision Workshops*, pp. 514-521, 2009.

[29] M. Jain, H. Jegou, and P. Bouthemy, "Better exploiting motion for better action recognition," *in Proceeding of IEEE Computer Vision and Pattern Recognition*, pp. 2555-2562, 2013

[30] S. Zhu, and L. Xia, "Human action recognition based on fusion features extraction of adaptive background subtraction and optical flow model," *Journal of Mathematical Problems in Engineering, Hindawi*, vol. 2015, 2015.

[31] D. Tsai, W. Chiu, and M. Lee, "Optical flow-motion history image (OF-MHI) for action recognition," *Journal of Signal, Image and Video Processing*, vol. 9, no. 8, pp. 1897-1906, 2015.

[32] L. Wang, Y. Qiao, and X. Tang, "MoFAP: A Multi-level Representation for Action Recognition," *International Journal of Computer Vision (IJCV)*, vol. 119, no. 3, pp. 254-271, 2016.

[33] W. Kim, J Lee, M. Kim ei al., "Human action recognition using ordinal measure of accumulated motion," *EURASIP Journal on Advances in Signal Processing, Hindawi*, vol. 2010, 2010.

[34] W. Zhang, Y. Zhang, C. Gao ei al., "Action recognition by joint spatial-temporal motion feature," *Journal of Applied Mathematics, Hindawi*, vol. 2013, 2013

[35] H. Tu, L. Xia, and Z. Wang, "The complex action recognition via the correlated topic model," *The Scientific World Journal, Hindawi*, vol. 2014, 2014.

[36] L. Gorelick, M. Blank, E. Shechtman et al., "Actions as Space-Time Shapes," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 12, pp. 2247-2253, 2007.

[37] A. Yilmaz and M. Shah, "A differential geometric approach to representing the human actions," *Journal of Computer Vision and Image Understanding*, vol. 109, no. 3, pp. 335-351, 2008.

[38] M. Grundmann, F. Meier, and I. Essa, "3D shape context and distance transform for action recognition," *in Proceeding of 19th International Conference on Pattern Recognition (ICPR)*, pp. 1-4, 2008.

[39] D. Batra, T. Chen, and R. Sukthankar, "Space-time Shapelets for action recognition," *in Proceeding of IEEE Workshop on Motion and video Computing(WMVC)*, pp. 1-6, 2008.

[40] S. Sadek, A. Al-Hamadi, G. Krell et al., "Affine-invariant feature extraction for activity recognition," *International Scholarly Research Notices on Machine Vision, Hindawi*, vol. 2013, 2013.

[41] C. Achard, X. Qu, A. Mokhber et al., "A novel approach for recognition of human actions with semi-global features," *Journal of Machine Vision and Applications*, vol. 19, no. 1, pp. 27-34, 2008.

[42] L. Dı´az-Ma´s, R. Mun˜oz-Salinas, F. J. Madrid-Cuevas et al., "Three-dimensional action recognition using volume integrals," *Journal of Pattern Analysis and Applications*, vol. 15, no. 3, pp. 289-298, 2012.

[43] K. Rapantzikos, Y. Avrithis, and Stefanos Kollias, "Spatiotemporal Features for Action Recognition and Salient Event Detection," *Journal of Cognitive Computation*, vol. 3, no. 1, pp. 167-184, 2011.

[44] N. Ikizler and P. Duygulu, "Histogram of oriented rectangles: A new pose descriptor for human action recognition," *Journal of Image and Vision Computing*, vol. 27, pp. 1515-1526, 2009.

[45] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," *in Proceeding of IEEE Computer Vision and Pattern Recognition*, pp. 1-8, 2008.

[46] V. Kellokumpu, G. Zhao and M. Pietikäinen, "Human activity recognition using a dynamic texture based method," *in Proceeding of the British Machine Vision Conference (BMVC'08)*, pp. 885-894, 2008.

[47] D. Tran, A. Sorokin, and D. A. Forsyth, "Human activity recognition with metric learning," *in Proceeding of the European Conference on Computer Vision (ECCV'08) - part 1*, pp. 548-561, 2008.

[48] B. Wang, Y. Liu, W. Wang et al., "Multi-scale locality-constrained spatiotemporal coding for local feature based human action recognition," *The Scientific World Journal, Hindawi*, vol. 2013, 2013.

[49] A. Gilbert, J. Illingworth, and R. Bowden, "Action recognition using mined hierarchical compound features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, 2011.

[50] I. C. Duta, J. R. R. Uijlings, B. Ionescu et al., "Efficient human action recognition using histograms of motion gradients and VLAD with descriptor shape information," *Journal of Multimedia Tools and Applications*, vol. 76, no. 21, pp. 22445-22472, 2016.

[51] M. Ahad, M. Islam, and I. Jahan, "Action recognition based on binary patterns of action-history and histogram of oriented gradient," *Journal on Multimodal User Interfaces*, vol. 10, no. 4, pp. 335-344, 2016.

[52] H. Rahmani and A. Mian, "3D action recognition from novel viewpoints," in Proceeding of IEEE Computer Vision Pattern Recognition, pp. 1506-1515, 2016.

[53] R. Kavi, V. Kulathumani, F. Rohit et al., "Multi-view fusion for activity recognition using deep neural networks," Journal of Electronic Imaging, vol. 25 no.4, 2016.

[54] Y. Kong, Z. Ding, Jun Li et al., "Deeply learned view-invariant features for cross-view action recognition," IEEE Transactions on image processing, vol. 26, no. 6, 2017.

[55] Mohamed A. Naiel, Moataz M. Abdelwahab and Motaz El-Saban, "Multi-view Human Action Recognition System Employing 2DPCA," In Applications of Computer Vision (WACV) IEEE Workshop, pp.270-275, 2011.

[56] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 23, no. 3, 2012.

[57] S. Pehlivan and P. Duygulu, "A new pose-based representation for recognizing actions from multiple cameras," *in Proceeding of Computer Vision and Image Understanding*, vol. 115, pp. 140-151, 2010.

[58] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," *in Proceeding of IEEE Computer Vision Pattern Recognition*, pp. 3209-3216, 2011.

[59] N. Gkalelis, N. Nikolaidis, and I. Pitas, "View independent human movement recognition from multi-view video exploiting a circular invariant posture representation," *in Proceeding of IEEE Multi Media and Expo*, pp. 394-397, 2009.

[60] R. Souvenir and J. Babbs, "Learning the viewpoint manifold for action recognition," *in Proceeding of IEEE Computer Vision Pattern Recognition*, pp. 1-7, 2008.

[61] M. Ahmad and S. W. Lee, "Hmm-based human action recognition using multi-view image sequences," *in Proceeding of IEEE Pattern Recognition*, pp. 263-266, 2006.

[62] A. Yao, J. Gall, and L. Gool, "Coupled action recognition and pose estimation from multiple views," *International Journal of Computer Vision (IJCV)*, vol. 100, no. 1, pp. 16-37, 2012.

[63] X. Ji, Z. Ju, C. Wang et al., "Multi-view transition HMMs based view-invariant human action recognition method," *Journal of Multimedia Tools and Applications*, vol. 75, no. 19, pp. 11847-11864, 2016.

[64] A. Kushwaha, S. Srivastava, and R. Srivastava, "Multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns," *Journal of Multimedia Systems*, vol. 23, no. 4, pp. 451-467, 2017.

[65] S. Spurlock, and R. Souvenir, "Dynamic view selection for multi-camera action recognition," *Journal of Machine Vision and Applications*, vol. 27, no.1, pp. 53-63, 2016.

[66] A. Chaaraoui, and F. Flórez-Revuelta, "A low-dimensional radial silhouette-based feature for fast human action recognition fusing multiple views," *International Scholarly Research Notices, Hindawi*, vol. 2014, 2014.

[67] A. Iosifidis, A. Tefas, and I. Pitas, "View-independent human action recognition based on multi-view action images and discriminant learning," *in proceeding IEEE Image, Video, and Multidimensional Signal Processing Workshop 2013*, pp.1-4, 2013.

[68] A. Liu, N. Xu, Y. Su et al., "Single/multi-view human action recognition via regularized multi-task learning," *Journal of Neurocomputing*, vol. 151, pp. 544-553, 2015.

[69] J. Wang, X. Nie, Yin Xia et al., "Cross-view action modeling, learning, and recognition," *in Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2649-2656, 2014.

[70] P. Huang, A. Hilton and J. Starck, "Shape similarity for 3D video sequences of people," *Journal of Computer Vision Springer*, vol. 89, no. 2, pp. 362-381, 2010.

[71] A. Veeraraghavan, A. Srivastava, A. Roy-Chowdhury et al., "Rate-invariant recognition of humans and their activities," *IEEE Transactions on Image Processing*, vol. 18, no. 6, 2009.

[72] I. Junejo, E. Dexter, I. Laptev et al., "View-independent action recognition from temporal self-similarities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, 2011.

[73] A. Iosifidis, A. Tefas, N. Nikolaidis et al., "Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis," *Journal of Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 347-360, 2012.

[74] N. Noorit, N. Suvonvorn, and M. Karnchanadecha, "Model-based human action recognition," *in Proceeding of Digital Image Processing SPIE*, 2010.

[75] M. Ahmad and S. Lee, "Human action recognition using shape and CLG-motion flowfrom multi-viewimage sequences," *Journal of Pattern Recognition*, vol. 41, pp.2237-2252, 2008.

[76] C. H. Chuang, J.W. Hsieh, L. W. Tsai et al., "Human action recognition using star templates and delaunay triangulation," *in Proceeding of Intelligent Information Hiding and Multimedia Signal Processing*, pp. 179-182, 2008.

[77] G. I. Parisi, C. Weber, S. Wermter, "Human action recognition with hierarchical growing neural gas learning," *in Proceeding of Artificial Neural Networks and Machine Learning on Springer*, pp. 89-96, 2014.

[78] N. Sawant and K. K. Biswas, "Human action recognition based on spatio-temporal features," *in Proceeding of Pattern Recognition and Machine Intelligence Springer*, pp. 357-362, 2009.

[79] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera Human Tracking with a Probabilistic Occupancy Map," *In IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol 30, no. 2, pp. 267-282, 2008.

[80] T. Zhao, M. Aggarwal, R. Kumar and H. Sawhney, "Real-time Wide Area Multi-Camera Stereo Tracking," *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 976-983, 2005.

[81] A. Mittal and L. S. Davis, "M_2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene," *International Journal of Computer Vision*, vol. 51, no.3, pp. 189-203, 2003.

[82] R. Muñoz-Salinas, R. Medina-Carnicer, F.J. Madrid-Cuevas, A. Carmona-Poyato, "Multi-camera human tracking using evidential filters" *In International Journal of Approximate Reasoning*, vol. 50, pp. 732 - 749, 2009.

[83] W. Hu, Min Hu, Xue Zhou, Tieniu Tan, Jianguang Lou and Maybank S., "Principal Axis-Based Correspondence between Multiple Cameras for Human Tracking," *In Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol 28, pp. 663 - 671, 2006.

[84] T.Chang and S. Gong, "Tracking multiple human with a multi-camera system," *In Multi-object tracking IEEE Conference*, pp. 19 - 26, 2001.

[85] S. Khan, O. Javed, Z. Rasheed, and Mubarak Shah, "Human Tracking in Multiple Cameras," *International Conference on Computer Vision*, 2001.

[86] R. Jain, R. Kasturi, and Brian G. Schunck, "Machine Vision," *Published by McGraw-Hill, Inc.*, ISBN 0-07-032018-7, 1995.

[87] R.C. Gonzalez, and R.E. Woods, "Digital Image Processing," *Published by Pearson, Inc.*, ISBN-13 978-0133356724, 2017.

[88] C. Stauffer, W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," *in Proceedings of Computer Vision and Pattern Recognition IEEE* Comput. Soc. Part Vol. 2, 1999.

[89] C. Stauffer, W. E. L. Grimson, "Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 22, no. 8, 2000.

[90] P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," *in Proceeding of Advanced Video-Based Surveillance Systems Springer*, pp. 135-144, 2001.

[91] T. Horprasert , D. Harwood, and L.S.Davis "A statistical approach for real-time robust background subtraction and shadow detection," *in IEEE ICCV'99 Frame-Rate Workshop*, 1999.

[92] H. Freeman, "On the classification of line-drawing data," *Models for the Perception of Speech and Visual Form*, pp. 408-412, 1967.

[93] D. Douglas and T. Peucker, "Algorithms for the reduction of the number of points required for represent a digitized line or its caricature," *Canadian Cartographer*, Vol.10, pp.112-122, 1973.

[94] A.K. Jain, J. Mao, and K.M. Mohiuddin, "Aritificial Neural Networks: A Tutorial," *IEEE Journal & Magazines*, Vol.29 No.3, 1996.

[95] DTREG, https://www.dtreg.com/solution/view/20, (last modified: unknown, Access date: 10 November 2018).

[96] N. Suvonvorn, "Prince of Songkla University (PSU) Multi-view profile-based action RGB-D dataset," 2017, http://fivedots.coe.psu.ac.th/~kom/?p=1483 (Access date: 20 December 2017).

[97] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis and I. Pitas, "The i3DPost multi-view and 3D human action/interaction," *in Proceeding of visual media production*, pp. 159-168, 2009.

ภาคผนวก ก.
ผลงานตีพิมพ์เผยแพร่จากวิทยานิพนธ์ 1

1. Pongsagorn Chalearnnetkul and Nikom Suvonvorn, "*High Level Fusion of Profile-Based Human Action Recognition using Multi-View RGBD Information*," in *Proceedings of the International Joint Conference on Computer Science and Software Engineering, JCSSE 2015*, 2015, pp. 36-40.


JCSSE 2015
2015 12th International Joint Conference
on Computer Science and Software Engineering (JCSSE)
"Shaping the Future with Convergence"
22 – 24 July 2015
Prince of Songkla University, Hat Yai, Songkhla, Thailand
Workshop on e-Science and High Performance Computing (eHPC 2015)
IEEE Catalog Number : CFP1532P-USB
ISBN : 978-1-4799-1965-9
SiPA
IEEE
IEEE
THAILAND SECTION
Technical Co-Sponsor
ECTI
Association
SIFT

2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE)

High Level Fusion of Profile-based Human Action Recognition using Multi-view RGBD Information

Pongsagorn Chalearnnetkul
Department of Computer Engineering
Faculty of Engineering, Prince of Songkla University
Hatyai, Songkhla, Thailand 90112
pongsagorn.ch@gmail.com

Nikom Suvonvorn
Department of Computer Engineering
Faculty of Engineering, Prince of Songkla University
Hatyai, Songkhla, Thailand 90112
nikom.suvonvorn@gmail.com

***Abstract*—Human action recognition system is fundamental of human activity and behavior recognition, especially for video analysis technologies. In this paper, we introduce an improvement method for human action recognition proposed by P.Chawalitsittikul *et al.* The actions from RGBD multi-views, taken from cameras at different static-viewpoints in the overlapping Area of Interest, are fused at high-level decision. Our empirical fusion model is derived from performances of action recognition in various viewpoints: front, slant, side, back-slant, and back. The results shown that the fusion model improves significantly the accuracy using only one more camera.**

Keywords-Action recognition; multi-view; fusion; viewpoint indicator; RGBD

I. INTRODUCTION

Undesirable situation could happen all the time. In the 3th generation of security system, the intelligent video analytics will be employed for improving the daily life using different advanced methodologies. Eventually, the better understanding of human behavior, the system could provide the better quality of life. The human action recognition, such as standing, walking, sitting, bending and laying, is fundamental for understanding the complex human behavior. In our previous work, we focus on the human action analysis using only single-view at a fixed viewpoint. Its main limitation is that there is not enough information to cover perfectly the scene for solving some serious situations, such as occlusion, obstruction, lost information, and etc. In this paper, we propose a multi-view approach to solve the problems, so using multiple cameras at different static-viewpoints.

In human tracking, two main groups of researches for the multi-view approach have been adopted to solve these problems. The first group based on feature matching [2][3], e.g. color histogram, texture, object features, trajectories [4]. The second is relied on 3D information and alignment, e.g. epipolar geometry [5], 3D volume coordination and projection [6][7], field of views [8], ground plane constraints, and landmark modality. In general, the calibrated system provides more robustness and accuracy than the first approach, but come with difficulty and complexity.

In action recognition, the multi-view fusion is mostly based on feature level. Researches can be grouped into two major approaches: 3D and 2D. The 3D emphasizes on 3D reconstructed modeling for representing the features or descriptors of human from multi-views. The 3D approach is often reliable and efficient for calibrated system, such as fixed orientation and number of cameras, and it comes with time consumption and complexity. The examples of 3D model for human action classification: volumetric motion history [9], 3D optical flows from harmonic 3D motion fields and correlation [10], Gaussian mixture model from voxel images [11], skeletal and super-quadratic modeling using voxels [12], hierarchical and cylindrical human modeling [13], etc. In 2D approach, human modeling use various feature representation schemes from 2D multi-view images. The models are often adjustable and flexible for arbitrary parameters and calibration, and feasible for real-time recognition, e.g., simultaneous multiple views action recognition using R-transform [14], bag of visual word features [15], 2D motion history and dynamic Bayesian [16], dynamic scene geometry [17], etc.

In multi-view high-level fusion, action recognition is the combination of the weighting actions from several single-view into a reliable result. M.A. Naiel [18] proposed high level multi-view fusion using 2DPCA of motion history features in order to classify action in single-view, and used the majority votes with minimum distance to determine action result.

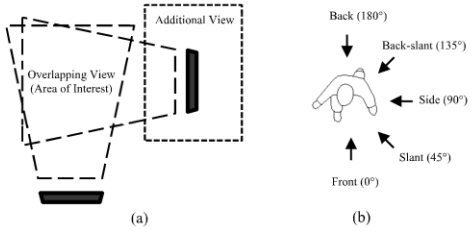


Figure 1. Multi-view setup (b) Viewpoint definitions.

In this paper, we introduce a high-level fusion method for improving the single-view human action recognition, proposed by P.Chawalitsittikul et al. [1]. The method uses the RGB and Depth information from multi-views taken from two cameras perpendicularly at different static-viewpoints in the overlapping Area of Interest, shown in figure 1(a).

978-1-4799-1966-6/15/$31.00 ©2015 IEEE

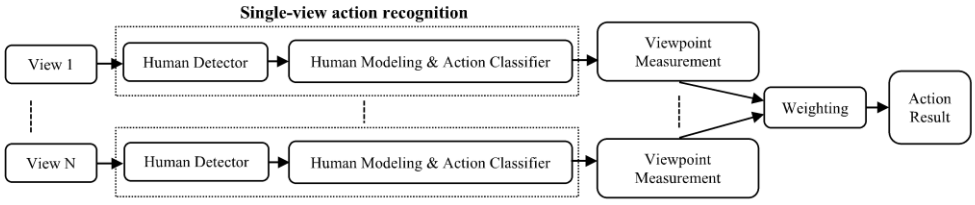


Figure 2. High Level Fusion of Profile-based Human Action Recognition using Multi-view system.

Our concept of high-level fusion technique is derived from the observation of action recognition in various viewpoints that its performances are vary with respect to the viewpoint. For example, laying action has high accuracy in the slant and side viewpoint, but get less rate in the front and back viewpoints. Our method will improve the accuracy by establishing the weighting function of action recognition rates with respect to the viewpoints, such as front, slant, side, back slant, and back, shown in figure 1(b).

II. Single-View Action Recognition

In this section, we describe the single-view of action recognition method proposed by P.Chawalitsittikul et al. Firstly, human detection is segmented from depth image using adaptive background subtraction method. Secondly, the head and legs position features of human model are determined. Finally, the artificial neural network is applied to the features for action classification into five basic actions.

A. Human Detection

The adaptive background subtraction, using Gaussian mixture-based background/foreground segmentation algorithm [19], is performed for extracting the motion that assumed to be human. This algorithm is robust to variant background which is defined by a mixture of 3-5 Gaussian distributions. The human object obtained from segmentation is then filled by depth information and its color edge. The color edge is the intersect of color image ($\boldsymbol{I_C}$) and object edge in depth image ($\boldsymbol{O_D} - E(\boldsymbol{O_D})$). The edge is defined by the subtraction of human depth object ($\boldsymbol{O_D}$) and erosion (E) of $\boldsymbol{O_D}$, descripted in equation (1).

$$\mathbf{H_{DC}} = \mathrm{E}(\mathbf{I_D}) \cup (\mathbf{I_C} \cap (\mathbf{O_D} - \mathrm{E}(\mathbf{O_D}))) \quad (1)$$

B. Human Modeling and Classification Method

In human modeling, the process starts with finding the center of mass in human object that refer to the head and legs positions. Then, the vectors with the maximum magnitude, from center of mass to human edge in each quadrant, are determined. The four vectors will be collapsed into two vectors, representing the head and legs respectively, by considering the least angle between any two vectors. The collapsing of vector is weighted by magnitude (D_v) and color distance (D_c) of vectors derived from equation (2) and (3). The color distance for the next frame (Cr_{t+1}) can be updated from color in last frame (Cr_t) by equation (4). The weighting value of first vector (ω_1) for combining is defined in equations (5) and (6).

$$D_{v_i} = \sqrt{(\Delta_{x_i} - \bar{x})^2 + (\Delta_{y_i} - \bar{y})^2} \mid i = 1 \ldots 4 \quad (2)$$

$$D_{c_i} = \sqrt{(\Delta_{c_i} - Cr_t)^2} \mid i = 1 \ldots 4 \quad (3)$$

$$Cr_{t+1} = \alpha_c Cr_t - (1 - \alpha_c)(\omega_1 C_1 + \omega_2 C_2) \quad (4)$$

$$\omega_{v_1} = \left[\sum_{j=1}^{2} \left(\frac{D_{v_2}}{D_{v_j}}\right)^{\frac{2}{m-1}}\right]^{-1} \quad (5)$$

$$\omega_{c_1} = \left[\sum_{j=1}^{2} \left(\frac{D_{c_2}}{D_{c_j}}\right)^{\frac{2}{m-1}}\right]^{-1} \quad (6)$$

The weighting value will be adjusted by magnitude (D_v) and color distance (D_c) using alpha ratio (α_v), shown in equation (7). Then, the second vector (ω_2) is simply defined as a complement of ω_1 using equation (8).

$$\omega_1 = \alpha_v \omega_v + (1 - \alpha_v)\omega_c \quad (7)$$

$$\omega_2 = 1 - \omega_1 \quad (8)$$

Finally, the positions of two vectors pointed to head and legs are combined into a single vector using weights of each vector following equation (9).

$$P(x, y) = \omega_1 P_1(\Delta_{x_1}, \Delta_{y_1}) + \omega_2 P_2(\Delta_{x_2}, \Delta_{y_2}) \quad (9)$$

The features of human model, which are defined using upper vector (center to head $\vec{v}_h$) and lower vector (center to legs $\vec{v}_l$), consist of the angle (θ_h) between body and head $\vec{v}_h$, the angle (θ_l) between body and legs $\vec{v}_l$, the disparity (D_h) for the difference of depth value between head and center, and the disparity (D_l) for the difference of depth value between legs and center. These features are necessary for classifying the actions. The artificial network is used as classification method, which is trained by back-propagation using single hidden layer.

III. High Level Multi-view Action Fusion

In our high-level fusion concept, the method is based on a weighting technique for determining the best action from different viewpoints. The action with highest weight is preferred that would increase the accuracy of recognition rate. The establishment of method is sharped from the different empirical observations of the single-view action recognition in various viewpoints, as shown in figure 2.

A. Empirical Observation for Viewpoint Action Recognition

During our experimentation, we found that the accuracy of action recognition method depends on its viewpoints that human features obtained from each viewpoint are variance due to noises, perspective distortion, or lack of information. For example, the sitting action detected from the front viewpoint is less accurate than side viewpoint. Bending action recognized from the front, back-slant and back viewpoints is less accurate than from the side viewpoint. Hence, we experiment the recognition method, described in section II, in five viewpoints to find the accuracy of every actions. The results are shown in table I. The table II shows the confusion matrix of single-view action recognition. These results will be used further in the fusion process.

TABLE I. ACTION RECOGNITION ACCURACY WITH RESPECT TO VIEWPOINT

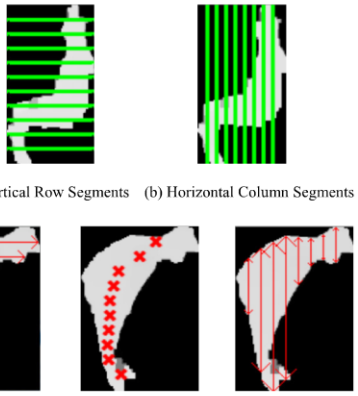
Desired Action	Viewpoint Accuracy (%)					
	Front	*Slant*	*Side*	*Back-slant*	*Back*	*Average*
Standing & walking	81.21	76.59	75.93	72.86	61.83	73.68
Sitting	85.61	44.75	53.72	29.97	93.00	61.41
Bending	62.57	73.61	85.80	3.95	6.32	46.45
Laying	0	78.67	98.06	89.56	0	53.25

TABLE II. CONFUSION MATRIX OF ACTION RECOGNITION ACCURACY

		Target Actions (%)			
		Standing & walking	*Sitting*	*Bending*	*Laying*
Desired Actions (%)	***Standing & walking***	74.14	23.75	32.27	27.06
	Sitting	23.22	61.90	18.69	10.33
	Bending	2.46	14.03	48.52	6.38
	Laying	0.18	0.32	0.52	43.35

B. Viewpoint Measurement

The measurement of viewpoint is the way to determine the viewpoint which is the direction of camera pointed to the observing human in action. The considering viewpoints consist of the front, slant, side, back-slant and back viewpoints, respectively. This measurement is very important in order to determine the accuracy of action recognition related to the viewpoint. The first process of feature determination is segmentation. The viewpoint is recognized from human features: width, axis and height. We define the positions of segment in order to calculate the features along human object vertically and horizontally following figure 3(a) and figure 3(b). For the width and axis features (shown in figure 3(c)), the extracted object of human is divided equally into ten segments vertically along the body from top to bottom, using equation (10). For each segment of ten width features ($WF(i)$), the width value is determined by the projection of white intensity pixel (I=1) in each row segment as described in equation 11. And axis feature ($AF(i)$) of any segments can be computed by finding position of first white pixel (LP_x) and last white pixel (RP_x) to define its center as equation 12.



(a) Vertical Row Segments (b) Horizontal Column Segments

(c) 10 Width, 10 Axis and 8 Height Features

Figure 3. Viewpoint Measurement Features,
(a) and (b) are vertical and horizontal segmentation in side sitting,
(c) is viewpoint feature in side bending consisted: 10 width features, 10 axis features and 8 height features.

The height features (shown in figure 3(c)) are divided equally into ten column segments along horizontal of human body from left to right following equation 13. The value of height can be calculated by projection of white intensity pixel (I=1) in the each column segment as described in equation 11.

$$PyWF(i), PyAF(i) = (\frac{\text{rows}}{11} * i) \mid i = 1 \dots 10 \quad (10)$$

$$WF(i) = \sum_{x=1}^{\text{cols}} f(x, PyWF(i)) \text{ If (I=1)} \quad (11)$$

$$AF(i) = LP_x(PyAF(i)) + (\frac{RP_x(PyAF(i)) - LP_x(PyAF(i))}{2}) \quad (12)$$

$$Px(i) = (\frac{\text{cols}}{9} * i) \mid i = 1 \dots 8 \quad (13)$$

$$HF(i) = \sum_{y=1}^{\text{rows}} f(\text{PxHF}_i, y) \text{ If (I=1)} \quad (14)$$

Finally, the artificial neural network is applied in order to classify the 5 types of viewpoint by using 28 human features as input.

C. Fusion Model

In this step, the fusion model is described. The model is established simply as weighting function. It is defined by the weight of accuracy rate of actions from every viewpoints based on the empirical learning result, shown in table I and table II.

The accuracy values of actions (A_{ij}) with corresponding to the specific viewpoint, depicted in table I, is determined as lookup table style, shown in equations (15).

$$A_{a_i,v_k} = \text{ta}[v_k][a_i] \quad (15)$$

Where, ta is table of actions with respect to viewpoint, v_k is the viewpoint k, and a_i is the action i.

For simple fusion model, we use the A_{a_i,v_k} value to choose the optimal action, e.g., if the sitting and standing actions are the results of the front and side viewpoints, which A values are 85.61 and 75.93. So, the sitting action is selected. However, from experimentation, we found that only this consideration is not enough to obtain the best result. So, we introduce more information to the decision is that the confusion matrix of actions.

The complex fusion model is defined by some additional parameters as following. The accuracy values of confusion actions ($C_{i,i}$) from every viewpoints, shown in table II, need to be determined as lookup table style, shown in equations (16)

$$C_{a_i^t,a_i^d} = \mathrm{tc}[a_i^t][a_i^d] \tag{16}$$

Where, tc is table of confusion matrix of actions, a_i^t and a_i^d are the target and desired action i correspondingly.

We establish a fusion function $F_{a_i a_j v_k}$ for an action a_i at viewpoint v_k with a penalty action a_j as follow:

$$F_{a_i a_j, v_k} = (C_{a_i^t,a_i^d} - \alpha A_{a_i,v_k}) - C_{a_i^t,a_j^d} \tag{17}$$

Where α is an adjustable parameter of our method.

For example, if the sitting and standing actions are the results of the front and side viewpoints, and the penalty actions are standing and sitting actions respectively. So, we need to compute the fusion function for both viewpoints separately

First viewpoint, the F value is weighted for sitting action as:

$$F_{a_{sitting} a_{stand}, v_{front}} = (61.90 - \alpha 85.61) - 23.75$$

And the second viewpoint, F value is weighted for standing action as:

$$F_{a_{stand} a_{sitting}, v_{side}} = (74.14 - \alpha 75.93) - 23.22$$

Then, we compare the F values for action selection. An action is chosen if only if F value from that viewpoint is maximum value.

IV. Experimental Results

Our system is implemented on Intel Core i7-4700MQ 2.4GHz using OpenCV and CLNUI libraries, acquired 640x480 pixels of RGBD information at 9 fps. Every single-view is processed on parallel using OpenMP. We test system with video datasets along 8700 frames for every action of two viewpoints, in which its directions are perpendicular, and focus on the overlapping area of interest. The range of processing is about 3 to 5.5 meters.

The time consumption is 110 ms for single-view and 240 ms for multi-view without parallel, and 180 ms for multi-view with parallel. Two experiments are performed: simple fusion model and proposed fusion model.

A. Simple Fusion Model

The experiment result is shown in table III. We can noticed that the simple fusion model can improve the majority of actions up to 98.71% for side laying, but dissatisfy for sitting. The false positive of sitting to be walking is about 23.75%. The overall results are increased at average 11.86 %. The experimental examples for this model can be shown in figure 4 (a-b).

TABLE III. Results from simple fusion model

Action	Viewpoint Accuracy (%)					
	Front	*Slant*	*Side*	*Back-slant*	*Back*	*Average*
Standing & walking	83.54	79.90	87.97	84.17	67.98	80.71
Sitting	68.87	26.10	68.58	17.67	88.48	53.94
Bending	82.34	60.05	84.66	59.32	39.12	65.10
Laying	77.66	96.89	98.71	94.95	44.22	82.49

B. Complex Fusion Model

We experiment our complex fusion model using $\alpha = 0.4$. The result is show in table IV. We can observed that the complex fusion model can significantly increase the accuracies of front bending and side laying actions up to 95.80% and 99.03%, respectively. However, the sitting and standing/walking are still but acceptable. The inferior result occurred on standing/walking while the complex model improves other actions because standing/walking is majority false desirable result of other actions. The overall results are increased at average 13.94 %. The experimental examples of complex fusion models show in figure 4(c-d).

TABLE IV. Result for complex fusion model

Action	Viewpoint Accuracy (%)					
	Front	*Slant*	*Side*	*Back-slant*	*Back*	*Average*
Standing & walking	51.44	58.95	68.15	70.48	55.99	61.00
Sitting	68.87	34.92	68.58	54.89	74.07	60.27
Bending	95.81	76.13	92.90	75.42	51.02	78.26
Laying	96.70	98.22	99.03	97.98	86.14	95.62

We found that some desirable results from our proposed fusion model is mainly caused by errors of the wrong prediction from the viewpoint measurement step, noted that its accuracy is only 84.9%

V. Conclusion And Further Work

In this paper, we proposed a high-level based action fusion method using multi-view to improve action recognition rate. The overall result rise up to 11.86 % and 13.94 % by using our proposed simple and complex fusion models. The maximum improvement for some actions is up to 96.70 %. In further work, the viewpoint measurement technique needs to be enhanced that might increase considerably the global result. Additionally, the feature levels of action fusion should be explored using the 3D reconstruction.

References

[1] P.Chawalitsittikul and N. Suvonvorn, "Profile-based Human Action Recognition using Depth Information," in proceedings of the IASTED International Conference on Advances in Computer Science and Engineering, ACSE 2012, 2012, pp. 376–380.

[2] K. Kyungnam, and L. S. Davis, "Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering," Computer Vision–ECCV 2006. Springer Berlin Heidelberg, 2006. 98-109.

[3] A. MITTAL and L. S. DAVIS, "M2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene," International Journal of Computer Vision, vol. 51(3), p.189-203, 2003

[4] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera Human Tracking with a Probabilistic Occupancy Map," In IEEE Trans on PAMI, Vol 30, NO. 2, pp 267-282, 2008

[5] T. H. Chang and S. Gong, "Tracking multiple human with a multi-camera system," In Multi-object tracking IEEE Conference, pp 19 – 26, 2001

[6] Y. Jian, and Jean-Marc Odobez, "Multi-camera multi-person 3D space tracking with MCMC in surveillance scenarios," Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2, 2008.

[7] R. Muñoz-Salinas, R. Medina-Carnicer, F.J. Madrid-Cuevas, A. Carmona-Poyato, "Multi-camera human tracking using evidential filters" In International Journal of Approximate Reasoning, Vol 50, pp 732 – 749, 2009

[8] S. Khan, O. Javed, Z. Rasheed, M. Shah, "Human tracking in multiple cameras," Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on. Vol. 1. IEEE, 2001.

[9] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," CVIU, vol. 104, no.2, pp. 249–257, 2006.

[10] M. Holte, T. Moeslund, N. Nikolaidis, and I. Pitas. "3d human action recognition for multi-view camera systems," In 3DIMPVT, 2011.

[11] S. Y. Cheng and M. M. Trivedi, "Articulated human body pose inference from voxel data using a kinematically constrained gaussian mixture model," CVPR Workshops, 2007.

[12] C. Tran and M. M. Trivedi, "Human body modeling and tracking using volumetric representation: Selected recent studies and possibilities for extensions," ACM/IEEE ICDSC, 2008.

[13] I. Mikic, M. M. Trivedi, E. Hunter, and P. Cosman, Human body model acquisition and tracking using voxel data, IJCV, vol. 53, no.3, pp. 199–223, 2003.

[14] R. Souvenir and J. Babbs, "Learning the viewpoint manifold for action recognition," In CVPR, 2008.

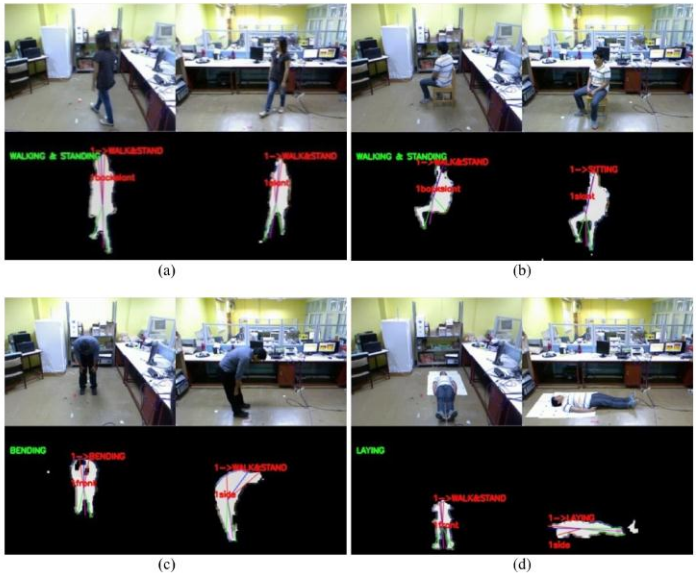
[15] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," In CVPR, 2011.

[16] S. Vitaladevuni, V. Kellokumpu, and L. Davis, "Action recognition using ballistic dynamics," In CVPR, 2008.

[17] A. Haq, I. Gondal, and M, "Murshed. On dynamic scene geometry for view-invariant action matching," In CVPR, 2011.

[18] Mohamed A. Naiel, Moataz M. Abdelwahab and Motaz El-Saban, "Multi-view Human Action Recognition System Employing 2DPCA," In Applications of Computer Vision (WACV) IEEE Workshop, pp.270-275, 2011.

[19] P. KadewTraKuPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," Proc. 2nd European Workshop on Advanced Video-Based Surveillance Systems, 2001



(a) (b) (c) (d)

Figure 4. Examples result: (a-b) simple model (c-d) complex model.

ภาคผนวก ข.
ผลงานตีพิมพ์เผยแพร่จากวิทยานิพนธ์ 2

2. Pongsagorn Chalearnnetkul and Nikom Suvonvorn, "A Rectangular Layer Model for Profile-Based Human Action Recognition using Multi-view Depth Information," in *Asia-Pacific Journal of Science and Technology (APST)*, Vol. 22, No. 3, Sep 30, 2017.


ISSN: 2539 - 6293
Asia - Pacific Journal of Science and Technology
APST

Asia-Pacific Journal of Science and Technology: Volume: 22. Issue: 03. Article ID.: APST-22-03-08. Research Article

APST

Asia-Pacific Journal of Science and Technology

https://www.tci-thaijo.org/index.php/APST/index

Published by the Research and Technology Transfer Affairs Division,
Khon Kaen University, Thailand

Rectangular Layer Model for Profile-based Human Action Recognition Using Multi-view Depth Information

Pongsagorn Chalearnnetkul[1]*, Nikom Suvonvorn[1]

[1]Department of Computer Engineering, Faculty of Engineering, Prince of Songkla University
*Correspondent author: pongsagorn.ch@gmail.com

Received June 2016
Accepted August 2016

Abstract

Human action recognition is a fundamental step for understanding complex activities or behaviors, especially for video surveillance and health-care applications. In this paper we introduce a profile-based human action recognition from multi-view cameras using RGB-D information through the Rectangular Layer Model (RLM). Our model tends to improve the performance when the perspective distortion or the lack of information occurs due to the single-view approach. The fusion model is tested for five basic actions: walking and standing, sitting, bending, and laying, at different perspective viewpoints. The system can perform at 28.99 fps while its overall precision is significant at about 92.25%.

Keywords: action recognition, multi-view, depth information, rectangular layer model

1. Introduction

Nowadays, intelligent video analytics play an important role in daily life for the detection and investigation of the abnormal events, especially for the security and health-care applications that could save lifes and properties. In the understanding of human behavior, the action analysis is a fundamental step to serves these goals. However, most research works on action analysis are still used information from only one viewpoint that does not allow to solve completely some serious problems such as occlusion, obstacle and lack of information. These problems can be solved intuitively by a multi-view approach using information from different viewpoints of the same Area of Interest. Nevertheless, a fusion technique is additionally needed, which is the key feature for combining information from the multi-view cameras. Recently, many interesting fusion methods are proposed, concerning either 2D or 3D techniques. The examples for 2D are follows: the bag of visual-words using spatio-temporal interest point for human modeling and classification [1], invariance multi-view action masks and movement representation vector [2], R-transform feature from simultaneous multiple views [3], silhouette feature space with PCA [4], etc. Concerning the 3D approach, the 3D human model will be reconstructed by fusing the features between views, where the camera calibration between viewpoints are needed that may increase the complexity and time consumption. The examples are follows: tracking of 3D joint of skeleton [5], human modeling with 3D circular volume [6], 3D optical flows in 3D motion context [7], and temporal shape similarity in 3D video [8]. In our previous works we focus mainly on 2D-3D techniques, which emphasize on the single-viewpoint of profile-based human action recognition using the 3D vector modeling [9], the interior 2D part movement [10], the deformable triangulation for skeleton extraction using string matching [11], the motion feature of depth map with hierarchical growing neural gas learning (GNG) [12] and the local histogram of optical-flow descripted by rectangular using Adaboost classifier [13].

In this paper we introduce a method for human action recognition from multi-view cameras using a feature-level fusion technique, called the Rectangular Layer Model (RLM). The RGB-D information is obtained from two cameras at different viewpoints, observing the specific Area of Interest. The RLM will fusion the features from these views to construct a human model, and classify the actions using

conventional classification methods. The system overview is illustrated in Figure 1.

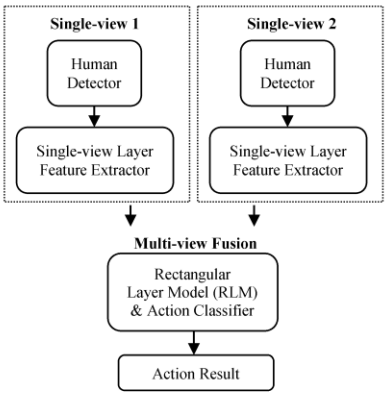


Figure 1. Overview of the human action recognition system using the Rectangular Layer Model.

2. Materials and Methods

2.1 Single-view processing

Single-view processing consists of two main functions: human detection and feature extraction. In preprocessing and human detection, it consists of five steps: motion detection, noise reduction, filling depth in motion object, object localization, and arm rejection. For motion detection, an adaptive background subtraction, the Gaussian mixture-based background/foreground segmentation algorithm [14], is applied in order to separate the human object from the depth background. The detected movement, or motion object, will be initialed as human. After that, the noise of motion object is reduced by using opening and closing morphological operators. This is followed by filling the depth information using AND operation, between motion object and depth information, as the motion-depth object. Then, we trace the boundary of the object using contour finding to locate the minimum rectangle or bounding box of the object that is used later in feature extraction. Additionally, a simple technique for arm rejection is applied in order to decrease the false alarm caused by the protruded parts that lead to false classification. The technique is based on the horizontal and vertical image projection, adapted from character segmentation algorithm, which is well-known in Optical Character Recognition (OCR).

In feature extraction step, each motion-depth object will be proceeded. We divided the features into two types: regular and penalty features. The regular features are represented in a layer format (Figure 2). Its concept is derived from the fact that the specific horizontal portions of human object, also displayed as layers, can represent the characteristic of different actions while allowing to reduce the dimension of feature data.

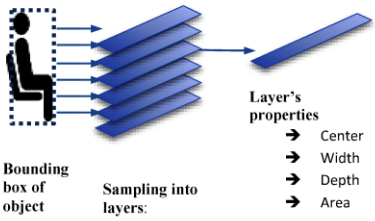


Figure 2. Rectangle layer features.

In the layer sampling step, the motion-depth object surrounded by the bounding box will be sampled horizontally $\mathcal{N}$ rows for $\mathcal{N}$ layers. A sampling row ($\mathcal{R}$) of layer could be demonstrated by equation (1). Where $\mathcal{r}$ is the number of object's rows and $\mathcal{N}$ is the number of layers.

$$\mathcal{R}(i) = \left(\frac{\mathcal{r}}{\mathcal{N}+1} \times i\right) \mid i = 1,2,\dots,\mathcal{N} \qquad (1)$$

After acquiring the row of each layer, its properties will be determined, for instance the width of layer ($\mathcal{W}$) and the depth of layer ($\mathcal{D}$). In each layer i, the $\mathcal{W}$ is calculated from the projection of bright-intensity pixel ($\mathcal{B}$) where its depth value is greater than a Depth Threshold ($\mathcal{T}$), as described in equation (2). The $\mathcal{D}$ is determined by the average of bright-intensity pixel ($\mathcal{B}$) of motion-depth object, defined by equation (3). Note that c the number of object's columns.

$$\mathcal{W}(i) = \sum_{x=1}^{c} \mathcal{B}\big(x, \mathcal{R}(i)\big) \qquad (2)$$

$$\mathcal{D}(i) = \frac{\sum_{x=1}^{c} f(\mathcal{B}(x, \mathcal{R}(i)))}{\sum_{x=1}^{c} \mathcal{B}(x, \mathcal{R}(i))} \qquad (3)$$

In order to correct the depth values which are non-linear according to the perspective distortion, depth of layer ($\mathcal{D}$) must be converted to real-depth ($\mathcal{D}_r$) in centimeter with respect to the intensity depth. We use the intensity-depth to real-depth conversion equation from P. Chawalitsittikul [15] that can be derived by polynomial regression power sixth as following equation (4).

$$\begin{aligned}\mathcal{D}_r = {} & 1.2512174874918e^{-10}\mathcal{D}^6 - \\ & 1.0370379397852e^{-7}\mathcal{D}^5 + \\ & 3.5014810721037e^{-5}\mathcal{D}^4 - \\ & 0.0061006393631\mathcal{D}^3 + \\ & 0.5775953878726\mathcal{D}^2 -\end{aligned} \qquad (4)$$

$27.6342681663553\mathcal{D} + 5.6759940397611e^2$

Finally, both properties are needed to be normalized using its max value, as show in equation (5) and (6).

$$\mathcal{W}_n(i) = \frac{\mathcal{W}(i)}{max(\mathcal{W})} \quad (5)$$

$$\mathcal{D}_n(i) = \frac{\mathcal{D}_r(i)}{max(\mathcal{D}_r)} \quad (6)$$

In addition, we establish a penalty feature that aims to support the classification of laying from other actions which is quite different. The feature is define as width-height ratio ($\propto$) of the bounding box of the object as follows equation (7):

$$\propto = \frac{c}{r} \quad (7)$$

The features mentioned above will be applied to the multi-view fusion using the models which will be detailed in the following section.

2.2 Multi-view feature fusion

In single-view, features might be insufficient perhaps to recognize actions affected by the perspective distortion and lack of information, thus the fusion of features from several single-views are necessary. Our proposed multi-view fusion method is based on the rectangular layer feature. Two models are established: Rectangular Layer Model (RLM) and Inverse Rectangular Layer Model (InvRLM).

Model 1 Rectangular Layer Model (RLM) is the simple model that will use only the width of layer ($\mathcal{W}_n$) to estimate the rectangular area feature ($\mathcal{RA}$) with the maximum of width-height ratio ($\propto_{max}$). The idea is demonstrated in Figure 3.

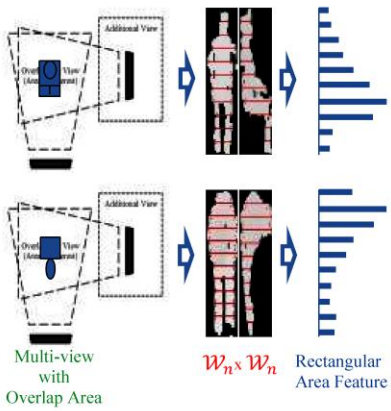


Figure 3. Demonstration of rectangular layer model in sitting and bending.

The Rectangular Area feature ($\mathcal{RA}$) in each layer is determined by the multiplication of the widths of layers ($\mathcal{W}_n$) from the corresponding layers, as shown in equation (8).

$$\mathcal{RA}(i) = \mathcal{W}_n(i)_{\text{view1}} \times \mathcal{W}_n(i)_{\text{view2}} \quad (8)$$

Then, it will be concatenated into a feature vector that is appropriate to training technique according to the specific classification method. The feature vector is defined as follows equation (9).

$$\vec{F}_{\mathcal{RA}} \in \{ \mathcal{RA}(1), \mathcal{RA}(2), \mathcal{RA}(3), \dots, \mathcal{RA}(\mathcal{N}), \propto_{max} \} \quad (9)$$

Model 2 Inverse Rectangular Layer Model (InvRLM) is derived from concept of the fact that when cameras are setup to view horizontally as overhead view manner, the depth value will inversely indicate the hidden volume of the object. For example, considering the sitting action from the front view, the depth value of the upper legs will be smaller than the other parts of body, while having the width of layer ($\mathcal{W}_n$) value nearly the other parts, as shown in Figure 4. Therefore, we established Inverse-Depth Weight Kernel ($\mathcal{IDK}$), as described in equation (10), which aims to adjust the rectangular area feature ($\mathcal{RA}$) to enhance the pattern of features that lead to a better classification in some actions. Thus, in the improved feature, the $\mathcal{W}_n$ will be multiplied by $\mathcal{IDK}$ in order to induce the corresponding actions, as a weighted $\mathcal{W}_n$ ($\mathcal{W}_w$), as shown in equation (11).

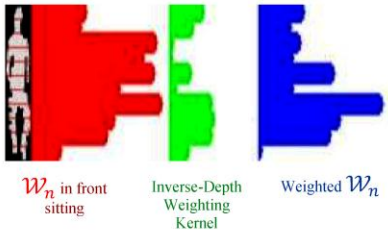


Figure 4. Example of Inverse Depth Weighting Kernel.

$$\mathcal{IDK}(i) = \frac{(\mathcal{D}_n(i) - max(\mathcal{D}_n))}{(min(\mathcal{D}_n) - max(\mathcal{D}_n))} \quad (10)$$

$$\mathcal{W}_w(i) = \mathcal{W}_n(i) \times \mathcal{IDK}(i) \quad (11)$$

However, we still believe in both width of layer ($\mathcal{W}_n$) and weighted width of layer ($\mathcal{W}_w$) features. Thus, the adjustment of the weight of these features is balanced by using $\tilde{\alpha}$ learning rate, which will provide the completely weighted width of layer ($\mathcal{W}_{\tilde{w}}$), as

shown in equation (12). Then, the weighted rectangular area feature ($\mathcal{RA}_w$) is estimated by equation (13). And finally, the $\mathcal{RA}_w$ will be concatenated with the maximum of width-height ratio ($\propto_{max}$) from all views into the feature vector is determined by equation (14), which is suitable for the classification.

$$\mathcal{W}_{\bowtie}(i) = ((1-\tilde{\alpha}) \times \mathcal{W}_n) + (\tilde{\alpha} \times \mathcal{W}_w) \quad (12)$$

$$\mathcal{RA}_w(i) = \mathcal{W}_{\bowtie}(i)_{\text{view1}} \times \mathcal{W}_{\bowtie}(i)_{\text{view2}} \quad (13)$$

$$\vec{F}_{\mathcal{RA}_w} \in \{\mathcal{RA}_w(1), \mathcal{RA}_w(2), \mathcal{RA}_w(3), \dots, \mathcal{RA}_w(\mathcal{N}), \propto_{max}\} \quad (14)$$

Action classification in our experiment of action classification, we use two traditional classification methods: artificial neural network (ANN) and support vector machine (SVM), applied to both fusion models. We trained the ANN with back-propagation algorithm using 10 and 20 hidden nodes for the Models 1 and 2 respectively. Also, we trained SVM using radial basis and C-SVC functions.

3. Results and Discussion

We notice that the standard benchmarks or datasets, focusing on actions analysis in depth data and multi-views, are not established. In experimentation, we then built our own datasets. The action range is about 3 to 5.5 meters from cameras that could acquire the full human body. The training dataset includes 8700 action frames in a clear room scene, whereas the angle between cameras are perpendicular. An example of the scene is shown in Figure 5. In addition, the testing dataset is built in different scene including 7800 action frames, with three different angles: 45°, 60° and 90° degrees. The example scene is shown in Figure 6. Our system is tested on CPU Intel Core i5 4590 at 3.30GHz using the OpenCV for computer vision library, OpenMP for parallel processing and CLNUI for acquiring the depth information from Kinect camera. We divide the single-view processing into threads which are simultaneously operated using parallel processing.

Figure 5. Example scene of experiment in trained dataset.

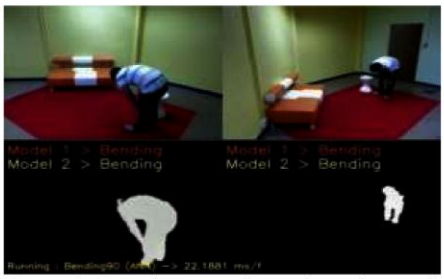


Figure 6. Example scene of experiment in tested dataset.

Regarding the time consumption, we measured the time using OpenMP wall-clock and OpenCV tick-frequency timer that does not affect different measurements of time consumption for models and classifiers. We found that the time consumption, or execution time, of each frame is around 35 milliseconds, or 29 fps, as shown in Table 1.

Table 1. Average time consumption of system.

Classifier	Time Consumption (ms)		Frame Rate (fps)	
	Model 1	Model 2	Model 1	Model 2
ANN	34.26	34.49	29.19	28.99
SVM	35.53	34.82	28.15	28.72

The average of time consumption in Table 1 shows that the time systems of different models are very similar. The minimum processing time is 34.26 ms in model 1 while using ANN Classifier. The processing frame rate is around 28-30 fps which coincided with the framerate of Kinect camera at 30 fps. As a result, the system can be performed in real-time.

The precision measurement is estimated from frames of both views where human are detected. The true positive is counted only if the true target actions are correctly classified, in contrast the false positive is when the recognized actions are incorrect. Though, the true and false negatives are not evaluated. To evaluate the fusion model, we defined a test dataset where the angle between cameras are different in order to test the robustness of models in every action. In addition, we tested the performance for both classification methods: artificial neural network and support vector machine, as shown in section A. and B. respectively. The testing dataset concerns the actions at normal speed and view of human in every perspective. The recorded videos of our experimental model using Artificial Neural Network (ANN) are available on https://www.youtube.com/watch?v=WRAq0A-0vhk (Evaluated on Test Dataset) and

https://www.youtube.com/watch?v=uTPiyOlnlFA (Evaluated on Trained Dataset).

3.1 Experimentation of RLM

We experimented the RLM using only rectangle areas as the least complex model. We tested both ANN and SVM, and the average of ANN result is 90.37% and 83.97% for SVM. It appears that ANN performs better than SVM following the Table 2. The majority of RLM results are significant with respect to actions and angles between cameras. However, the performance of sitting action is not adequate in the narrow angle, especially for the sitting in slant view; its features are similar to the Walking & Standing action. When the angle between the cameras is very narrow, the precision is reduced accordingly.

Table 2. Precision of RLM.

ACTION	ANGLE BETWEEN CAMERA					
	Precision of ANN (%)			*Precision of SVM (%)*		
	45°	*60°*	*90°*	*45°*	*60°*	*90°*
Walking & Standing	96.75	99.09	99.35	90.94	99.64	96.12
Sitting	69.17	75.90	87.67	50.78	52.87	63.85
Bending	92.35	87.16	83.76	88.71	86.53	80.13
Laying	98.82	97.47	96.97	98.63	100.00	99.42
Average in Angle	89.27	89.91	91.94	82.26	84.76	84.88
Overall		90.37			83.97	

3.2 Experimentation of InvRLM

The experimental results are detailed in Table 3. In summary, the overall accuracies of InvRLM model using ANN and SVM are 92.25% and 85.57% respectively. In this experiment, the weight adjustment parameter $\tilde{\alpha}$ are specified to 0.5 and 0.7. We found that the precision is 82.50% when $\tilde{\alpha}$ is 0.5 and 92.25% when $\tilde{\alpha}$ is 0.7. Therefore, the optimal value of $\tilde{\alpha}$ is 0.7.

We noticed that the InvRLM model, which is the enhancement of RLM by using properties of real-depth to weight the pattern of rectangle area, can improve the precision of sitting action which is the problem of RLM. However, it decreases the precision of laying action that is caused by the lack of body parts, which normally requires the full-part of human object. In most laying errors, the lack of human's body parts will cause the depth value to be close to background, which makes it very difficult to separate the motion.

Table 3. Precision of InvRLM.

ACTION	ANGLE BETWEEN CAMERA					
	Precision of ANN (%)			*Precision of SVM (%)*		
	45°	*60°*	*90°*	*45°*	*60°*	*90°*
Walking & Standing	95.90	98.18	93.53	91.11	95.09	84.48
Sitting	97.91	99.79	99.43	99.79	100.00	99.72
Bending	93.62	89.05	87.61	93.81	85.68	83.33
Laying	81.37	84.44	86.17	73.73	62.45	57.64
Average in Angle	92.20	92.86	91.68	89.61	85.81	81.29
Overall		92.25			85.57	

Table 4. Precision comparison.

Method	PRECISION RATES OF ACTION (%)					
	Standing	*Walking*	*Sitting*	*Bending*	*Laying*	*Average*
Our method	95.87	95.87	99.04	90.09	83.99	92.97
P. Chawalitsittikul [9]	98.00	98.00	93.00	94.10	98.00	96.22
N. Noorit [10]	99.41	80.65	89.26	94.35	100.0	92.73
Chi-Hung [11]	-	92.40	97.60	95.40	-	95.80
G. I. Parisi [12]	96.67	90.00	83.33	-	86.67	89.17
N. Sawant [13]	91.85	96.14	85.03	-	-	91.01

3.3 Precision Comparison

Table 4 shows the comparison results. We compare the performance of our methods with the single-view action recognition system, proposed by P. Chawalitsittikul [9] and N. Noorit [10]. The experiment of InvRLM is performed using ANN classifier (average values from all angles between cameras). We can notice that our work demonstrate approximately the same precision with the other methods. However, our method can perform good results in various orientations, and executed faster at 28.99 fps.

4. Conclusion

In this paper we proposed the multi-view fusion model for profile-based human action recognition using depth information. Our fusion models can achieve the action recognition at every camera viewpoint. The overall precision is at 92.25% in average by using InvRLM with ANN classifier, and can be performed in real-time at 28.99 fps.

In further works, we will apply new advanced methods for better action classification. In addition, we will consider the human model that can recognize more complex actions. Additionally, the features might be partially 3D for structural movement modeling.

5. Acknowledgement

This work was supported by Graduate School Thesis Grant, Prince of Songkla University 2015.

6. References

[1] Liu J, Shah M, Kuipers B, Savarese S. Cross-view action recognition via view knowledge transfer. Proc Computer Vision Pattern Recognition IEEE. 2011;3209-3216.

[2] Gkalelis N, Nikolaidis N, Pitas, I. View independent human movement recognition from multi-view video exploiting a circular invariant posture representation. Proc Multi Media and Expo IEEE. 2009;394-397.

[3] Souvenir R, Babbs J. Learning the viewpoint manifold for action recognition. Proc Computer Vision Pattern Recognition IEEE. 2008;1-7.

[4] Ahmad M, Lee SW. Hmm-based human action recognition using multi-view image sequences. Proc Pattern Recognition IEEE. 2006;1:263-266.

[5] Tran C, Trivedi MM. Human body modeling and tracking using volumetric representation Selected recent studies and possibilities for extensions. Proc Distributed Smart Cameras ACM/IEEE. 2008;1-9.

[6] Pehlivan S, Duyugulu P. A new pose-based representation for recognizing actions from multiple cameras. Journal of Computer Vision and Image Understanding ACM. 2010;115(2):140–151.

[7] Holte MB, Moeslund TB, Nikolaidis N, Pitas I. 3D human action recognition for multi-view camera systems. Proc 3D Imaging, Modeling, Processing, Visualization and Transmission IEEE. 2011;342-349.

[8] Huang P, Hilton A, Starck J. Shape similarity for 3D video sequences of people. Journal of Computer Vision Springer. 2010;89(2):362-381.

[9] Chawalitsittikul P, Suvonvorn N. Profile-based Human Action Recognition using Depth Information. Proc Advances Computer Science and Engineering ACTA Press. 2012;376–380.

[10] Noorit N, Suvonvorn N , Karnchanadecha M. Model-based Human Action Recognition. Proc Digital Image Processing SPIE. 2010;7546.

[11] Chuang CH, Hsieh JW, Tsai LW, Fan KC. Human Action Recognition Using Star Templates and Delaunay Triangulation. Proc Intelligent Information Hiding and Multimedia Signal Processing. 2008;179-182.

[12] Parisi GI, Weber C, Wermter S. Human action recognition with hierarchical growing neural gas learning. Proc Artificial Neural Networks Springer. 2014;8681:89–96.

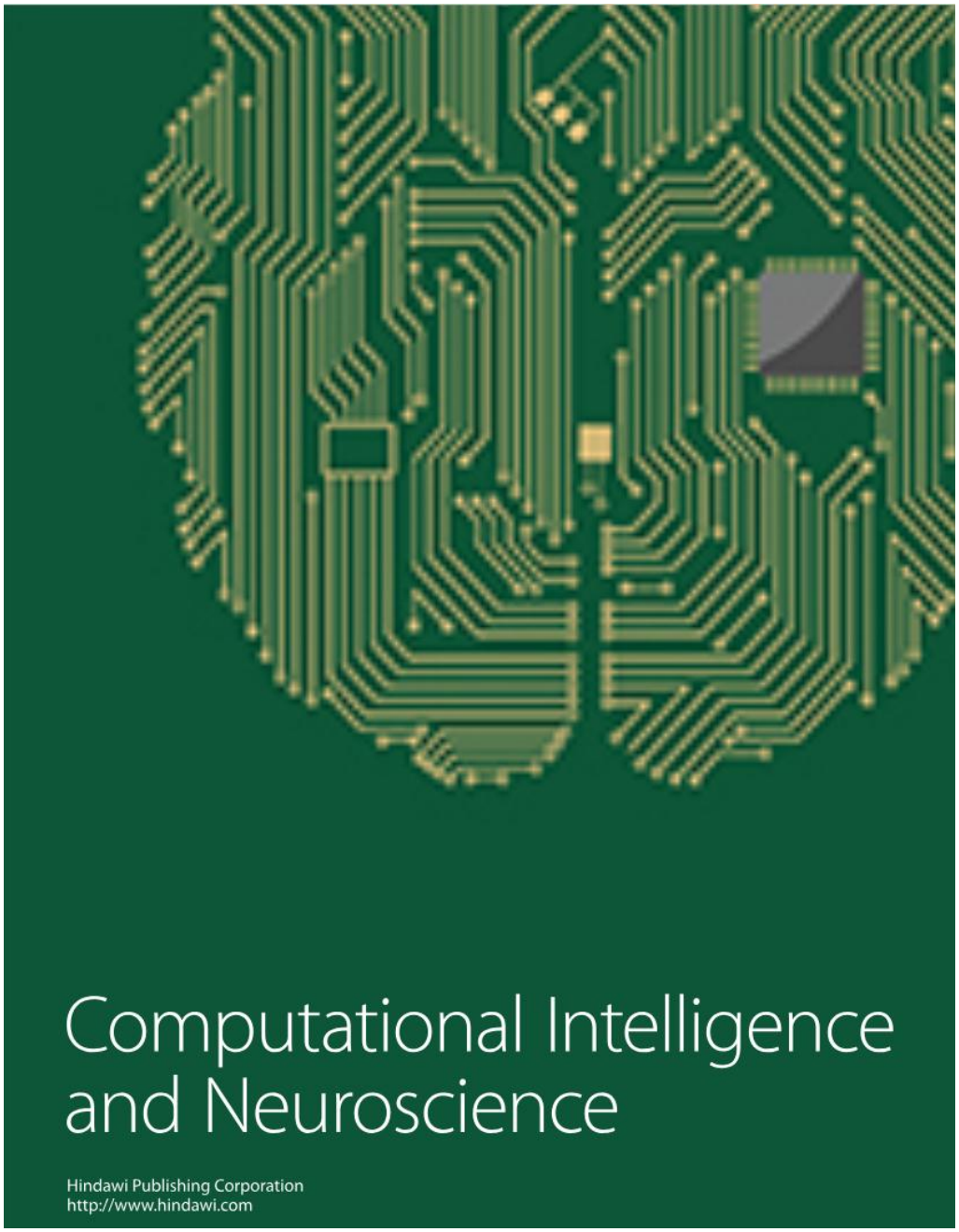
[13] Sawant N, Biswas KK. Human Action Recognition Based on Spatio-temporal Features. Proc Pattern Recognition and Machine Intelligence Springer. 2009;5909:357-362.

[14] KaewTraKuPong P, Bowden R. An improved adaptive background mixture model for real-time tracking with shadow detection. Proc Advanced Video-Based Surveillance Systems Springer, 2001;135-144.

[15] Chawalitsittikul P. Side-View Based Human Action Recognition Using Stereo Vision [Thesis]. Songkhla: Prince of Songkla University; 2013, Thai.

ภาคผนวก ค.
ผลงานตีพิมพ์เผยแพร่จากวิทยานิพนธ์ 3

3. Pongsagorn Chalearnnetkul and Nikom Suvonvorn, *"Multiview Layer Fusion Model for Action Recognition Using RGBD Images," Computational Intelligence and Neuroscience, vol. 2018, Article ID 9032945, 22 pages, 2018. https://doi.org/10.1155/2018/9032945.*


Computational Intelligence
and Neuroscience
Hindawi Publishing Corporation
http://www.hindawi.com

Hindawi
Computational Intelligence and Neuroscience
Volume 2018, Article ID 9032945, 22 pages
https://doi.org/10.1155/2018/9032945

Research Article

Multiview Layer Fusion Model for Action Recognition Using RGBD Images

Pongsagorn Chalearnnetkul and Nikom Suvonvorn

Department of Computer Engineering, Faculty of Engineering, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand

Correspondence should be addressed to Pongsagorn Chalearnnetkul; pongsagorn.ch@gmail.com

Received 3 January 2018; Revised 27 April 2018; Accepted 20 May 2018; Published 20 June 2018

Academic Editor: Pedro Antonio Gutierrez

Vision-based action recognition encounters different challenges in practice, including recognition of the subject from any viewpoint, processing of data in real time, and offering privacy in a real-world setting. Even recognizing profile-based human actions, a subset of vision-based action recognition, is a considerable challenge in computer vision which forms the basis for an understanding of complex actions, activities, and behaviors, especially in healthcare applications and video surveillance systems. Accordingly, we introduce a novel method to construct a *layer feature model* for a profile-based solution that allows the fusion of features for multiview depth images. This model enables recognition from several viewpoints with low complexity at a real-time running speed of 63 fps for four profile-based actions: standing/walking, sitting, stooping, and lying. The experiment using the Northwestern-UCLA 3D dataset resulted in an average precision of 86.40%. With the i3DPost dataset, the experiment achieved an average precision of 93.00%. With the PSU multiview profile-based action dataset, a new dataset for multiple viewpoints which provides profile-based action RGBD images built by our group, we achieved an average precision of 99.31%.

1. Introduction

Since 2010, action recognition methods have been increasingly developed and have been gradually introduced in healthcare applications, especially for monitoring the elderly. Action analysis plays an important role in the investigation of normal or abnormal events in daily-life activities. In such applications, privacy and convenience of usage of chosen technologies are two key factors that must be thoroughly considered. The pattern of recognized actions is an important function of a system for monitoring complex activities and behaviors which consist of several brief actions constituting a longer-term activity outcome. For example, a sleeping process involves standing/walking, sitting, and lying actions; and a falling process includes all actions mentioned above except sitting.

Recently, two main approaches have been studied and proposed for determining these actions: a wearable sensor-based technique and a vision-based technique.

Wearable inertial sensor-based devices have been used extensively in action recognition due to their small size, low power consumption, low cost, and the ease with which they can be embedded into other portable devices, such as mobile phones and smart watches. An inertial sensor used for performing navigation commonly comprises motion and rotation sensors, (e.g. accelerometers and gyroscopes). It provides the path of movement, viewpoint, velocity, and acceleration of the tracked subject. Some research studies have used wearable sensors [1–3], mobile phones [4–7], and smart watches [8] for recognizing different actions. In some research, the focus was on detection of abnormal actions, such as falling [9–11], or on reporting status for both normal and abnormal situations [12]. To recognize complex actions, moreover, several sensors must be embedded at different positions on the body. The only limitation of inertial sensors is the inconvenience presented because sensors must eventually be attached to the body, which is uncomfortable and cumbersome.

For vision-based techniques, many studies emphasize using either a single-view or multiview approach for recognizing human actions.

In a single-view approach, four types of feature representation have been used: (1) joint-based/skeleton-based, (2) motion/flow-based, (3) space-time volume-based, and (4) grid-based:

(1) Joint-based/skeleton-based representation defines the characteristics of human physical structure and distinguishes its actions, for example, multilevel of joints and parts from posing features [13], the Fisher vector using skeletal quads [14], spatial-temporal feature of joints-mHOG [15], Lie vector space from a 3D skeleton [16], invariant trajectory tracking using fifteen joints [17], histogram bag-of-skeleton-codewords [18], masked joint trajectories using 3D skeletons [19], posture features from 3D skeleton joints with SVM [20], and star skeletons using HMMs for missing observations [21]. These representations result in clear human modeling, although the complexity of joint/skeleton estimation requires good accuracy from tracking and prediction.

(2) Motion/flow-based representation is a global feature-based method using the motion or flow of an object, such as invariant motion history volume [22], local descriptors from optical-flow trajectories [23], KLT motion-based snippet trajectories [24], Divergence-Curl-Shear descriptors [25], hybrid features using contours and optical flow [26], motion history and optical-flow images [27], multilevel motion sets [28], projection of accumulated motion energy [29], pyramid of spatial-temporal motion descriptors [30], and motion and optical flow with Markov random fields for occlusion estimation [31]. These methods do not require accurate background subtractions but make use of acquired, inconstant features that need strategy and descriptors to manage.

(3) Volume-based representations are modeled by stacks of silhouettes, shapes, or surfaces that use several frames to build a model, such as space-time silhouettes from shape history volume [32], geometric properties from continuous volume [33], spatial-temporal shapes from 3D point clouds [34], spatial-temporal features of shapelets from 3D binary cube space-time [35], affine invariants with SVM [36], spatial-temporal micro volume using binary silhouettes [37], integral volume of visual-hull and motion history volume [38], and saliency volume from luminance, color, and orientation components [39]. These methods acquire a detailed model but must deal with high dimensions of features which require accurate human segmentation without the background.

(4) Grid-based representations divide the observation region of interest into cells, a grid, or overlapped blocks to encode local features, for example, a grid or histogram of oriented rectangles [40], flow descriptors from spatial-temporal small cells [41], histogram of local binary patterns from a spatial grid [42] and rectangular optical-flow grid [43], codeword features for histograms of oriented gradients and histograms of optical flow [44], 3D interest points within multisize windows [45], histogram of motion gradients [46], and combination of motion history, local binary pattern, and histogram of oriented gradients [47]. This method is simple for feature modeling in the spatial domain, but it must deal with some duplicate and insignificant features.

Although the four types of representation described in the single-view approach are generally good, in monitoring a large area, one single camera will lose its ability to determine continuous human daily-life actions due to view variance, occlusion, obstruction, and lost information, among others. Thus, a multiview approach is introduced to lessen the limitations of a single-view approach.

In the multiview approach, methods can be categorized into 2D and 3D methods.

Examples of the 2D methods are layer-based circular representation of human model structure [48], bag-of-visual-words using spatial-temporal interest points for human modeling and classification [49], view-invariant action masks and movement representation [50], R-transform features [51], silhouette feature space with PCA [52], low-level characteristics of human features [53], combination of optical-flow histograms and bag-of-interest-point-words using transition HMMs [54], contour-based and uniform local binary pattern with SVM [55], multifeatures with key poses learning [56], dimension-reduced silhouette contours [57], action map using linear discriminant analysis on multiview action images [58], posture prototype map using self-organizing map with voting function and Bayesian framework [59], multiview action learning using convolutional neural networks with long short term memory [60], and multiview action recognition with an autoencoder neural network for learning view-invariant features [61].

Examples of the 3D method, where the human model is reconstructed or modeled from features between views, are pyramid bag-of-spatial-temporal-descriptors and part-based features with induced multitask learning [62], spatial-temporal logical graphs with descriptor parts [63], temporal shape similarity in 3D video [64], circular FFT features from convex shapes [65], bag-of-multiple-temporal-self-similar-features [66], circular shift invariance of DFT from movement [67], and 3D full body/pose dictionary features with convolutional neural networks [68]. All of these 3D approaches attempt to construct a temporal-spatial data model that is able to increase the model precision and, consequently, raise the accuracy of the recognition rate.

The multiview approach, however, has some drawbacks. The methods need more cameras and hence are more costly. It is a more complex approach in terms of installation, camera calibration between viewpoints, and model building and hence is more time-consuming. In actual application, however, installation and setup should be simple, flexible, and as easy as possible. Systems that are calibration-free or automatically self-calibrating between viewpoints are sought.

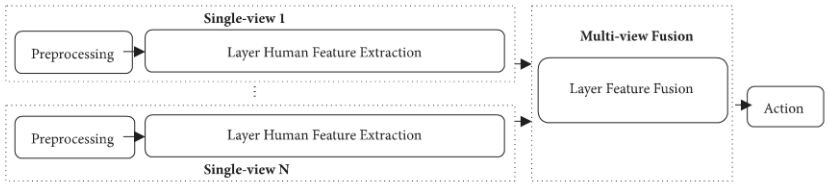


Figure 1: Overview of layer fusion model.

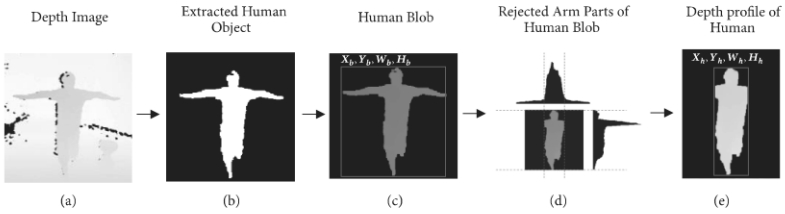


Figure 2: Preprocessing and human depth profile extraction; (a) 8-bit depth image acquired from depth camera; (b) motion detected output from mixture-based Gaussian model for background subtraction; (c) blob position: consists of top-left position (X_b, Y_b) and width (W_b) and height (H_b); (d) rejected arm parts of human blob; (e) depth profile and position of human blob: X_h, Y_h, W_h, and H_h.

One problem facing a person within camera view, be it one single camera or a multitude of cameras, is that of privacy and lighting conditions. Vision-based and profile-based techniques involve the use of either RGB or non-RGB. The former poses a serious problem to privacy. Monitoring actions in private areas using RGB cameras make those under surveillance feel uncomfortable because the images expose more clearly their physical outlines. As for lighting conditions, RGB is also susceptible to intensity; images often deteriorate in dim environments. The *depth approach* helps solve both problems; a coarse depth profile of the subject is adequate for determining actions, and depth information can prevent illumination change issues, which are a serious problem in real-life applications of round-the-clock surveillance. The depth approach that is adopted in our research together with a multiview arrangement is considered worthy of more costly installation than the single-view approach.

A gap that needs attention for most multiview, non-RGB results is that of perspective robustness, or viewing-orientation stability, and model complexity. Under a calibration-free setup, our research aims to contribute to the development of a fusion technique that is robust and simple in evaluating the depth profile of human action recognition. We have developed a *layer fusion model* in order to fuse depth profile features from multiviews and to test our technique on a triple dataset of validation and efficiency. The three datasets tested are the Northwestern-UCLA dataset, the i3DPost dataset, and the PSU dataset for multiview action from various viewpoints.

The following sections detail our model, its results, and comparisons.

2. Layer Fusion Model

Our *layer fusion model* is described in three parts: (1) preprocessing for image quality improvement; (2) human modeling and feature extraction using a single-view *layer feature extraction module*; and (3) fusion of features from any view into one single model using *layer feature fusion module* and classifying to actions. The system overview is shown in Figure 1.

2.1. Preprocessing. The objective of preprocessing is to segregate human structure from the background and to eliminate arm parts before extracting the features, as depicted in Figure 2. In our experiment, the structure of the human in the foreground is extracted from the background by applying motion detection using mixture of Gaussian model segmentation algorithms [69]. The extracted motion image ($\boldsymbol{I_m}$) (Figure 2(b)), from the depth image ($\boldsymbol{I_d}$) (Figure 2(a)), is assumed to be the human object. However,$\boldsymbol{I_m}$ still contains noise from motion detection, which has to be reduced by morphological noise removal operator. The depth of the extracted human object, defined by its depth values inside the object, is then improved ($\boldsymbol{I_{mo}}$) as determined by intersecting $\boldsymbol{I_d}$ and $\boldsymbol{I_m}$ using the AND operation: $\boldsymbol{I_{mo}} = \boldsymbol{I_m}\&\boldsymbol{I_d}$.

The human blob, with bounding rectangle coordinates X_h and Y_h, width W_h, and height H_h as shown in Figure 2(c), is located using a contour approximation technique. However, our action technique emphasizes only the profile structure of the figure, while hands and arms are excluded, as seen in Figure 2(d), and the obtained structure is defined as the figure depth profile ($\boldsymbol{I_{mo}}$) (Figure 2(e)) in further recognition steps.

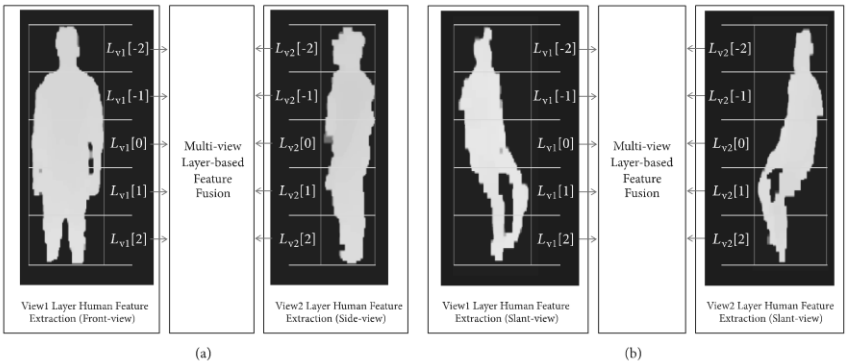


Figure 3: Layered human model for multiview fusion: (a) sampled layer model for standing; (b) sampled layer model for sitting.

2.2. Layer Human Feature Extraction. We model the depth profile of a human in a layered manner that allows extraction of specific features depending on the height level of the human structure, which possesses different physical characteristics. The depth profile is divided vertically into odd-numbered layers (e.g., 5 layers, as shown in Figure 3) with a specific size, regardless of the distance, perspective, and views, which would allow features of the same layer from all views to be fused into reliable features.

The human object in the bounding rectangle is divided into equal layers to represent features at different levels of the structure, $L[k] \mid k \in \{-N, -N+1, -N+2, \ldots, 0, 1, 2, N-2, N-1, N\}$. The total number of layers is $2N+1$, where N is the maximum number of upper or lower layers. For example, in Figure 3, the human structure is divided into five layers (N equals 2): two upper, two lower, and one center; thus the layers consist of $\{L[-2], L[-1], L[0], L[+1], \text{and } L[+2]\}$. The horizontal boundaries of all layers, the red vertical lines, are defined by the left and right boundaries of the human object. The vertical boundaries of each layer, shown as yellow horizontal lines, are defined by the top $y_T[k]$ and the bottom $y_B[k]$ values that can be computed as follows:

$$y_T[k] = \frac{H_h(k+N)}{2N+1} + 1 \tag{1}$$

$$y_B[k] = \frac{H_h(k+N+1)}{2N+1} \tag{2}$$

The region of interest in layer k is defined as $x = 0$ to W_h and $y = y_T[k]$ to $y_B[k]$.

According to the model, features from the depth profile human object can be computed along with layers as concatenated features of every segment using basic and statistical properties (e.g., axis, density, depth, width, and area). Depth can also be distinguished for specific characteristic of actions.

In our model, we define two main features for each layer, including the *density* ($\rho[k]$) and *weighted depth density* ($Z[k]$) of layers. In addition, the *proportion value* (P_v) is defined as a global feature to handle the horizontal action.

2.2.1. Density of Layer. The *density of layer* ($\rho[k]$) indicates the amount of object at a specific layer, which varies distinctively according to actions, and can be computed as the number of white pixels in the layer, as shown in the following equation:

$$\rho'[k] = \sum_{y=y_T[k]}^{y_h[k]} \sum_{x=0}^{W_h} I_m(x, y) \tag{3}$$

In multiple views, different distances between the object and the cameras would affect $\rho'[k]$. Objects close to a camera certainly appear larger than when further away, and thus $\rho'[k]$ must be normalized in order for these to be fused. We use the maximum value of the perceived object and normalize it, employing the following equation:

$$\rho[k] = \frac{\rho'[k]}{\arg\max(\rho'[k])} \tag{4}$$

2.2.2. Weighted Depth Density of Layer. An inverse depth density is additionally introduced to improve the pattern of density feature. The procedure is comprised of two parts: inverse depth extraction and weighting for density of the layer.

At the outset, depth extraction is applied to the layer profile. The depth profile reveals the surface of the object that indicates rough structure ranging from 0 to 255 or from near to far distances from the camera. According to perspective projection varying in a polynomial form, a depth value at a near distance, for example, from 4 to 5, has a much smaller real distance than a depth value at a far distance, for example, from 250 to 251. The real-range depth of the layer ($D'[k]$) translates the property of 2D depth values to real 3D depth values in centimeters. The real-range depth better distinguishes the depth between layers of the object—different

parts of the human body—and increases the ability to classify actions. A polynomial regression [70] has been used to convert the depth value to real depth, as described in the following equation:

$$f_c'(x) = 1.2512e^{-10}x^6 - 1.0370e^{-7}x^5 + 3.5014e^{-5}x^4 - 0.0061x^3 + 0.5775x^2 - 27.6342x + 5.6759e^2 \quad (5)$$

In addition, the $(D'[k])$ value of each layer is represented by the converted value of every depth value averaged in that layer, as defined in the following equation:

$$D'[k] = f_c'\left(\frac{\sum_{y=y_T[k]}^{y_B[k]}\sum_{x=0}^{W_h} I_{mo}(x,y)}{\rho[k]}\right) \quad (6)$$

Numerically, to be able to compare the depth profile of human structure from any point of view, the $D'[k]$ value needs to be normalized using its maximum value over all layers by employing the following equation:

$$D[k] = \frac{D'[k]}{\arg\max(D'[k])} \quad (7)$$

In the next step, we apply the inverse real-range depth $(D_i[k])$, hereafter referred to as the inverse depth, for weighting the density of the layer in order to enhance the feature $\rho[k]$ that increases the probability of classification for certain actions, such as sitting and stooping. We establish the inverse depth, as described in (8), to measure the hidden volume of body structure which distinguishes particular actions from others: for example, in Table 1, in viewing stooping from the front, the upper body is hidden but the value of $D_i[k]$ can reveal the volume of this zone; and in viewing sitting from the front, the depth of the thigh will reveal the hidden volume compared to other parts of the body.

$$D_i[k] = \left(\frac{D[k] - \arg\max(D[k])}{\arg\min(D[k]) - \arg\max(D[k])}\right) \quad (8)$$

The inverse depth density of layers $(Q[k])$ in (9) is defined as the product of the inverse depth $(D_i[k])$ and the density of the layer $(\rho[k])$.

$$Q[k] = (D_i[k])(\rho[k]) \quad (9)$$

We use learning rate α as an adjustable parameter that allows balance between $Q[k]$ and $\rho[k]$. In this configuration, when the pattern of normalized inverse depth $D_i[k]$ is close to zero, $Q[k]$ is close to $\rho[k]$. Equation (10) is the *weighted depth density of layers* $(Z[k])$, adjusted by the learning rate on $Q[k]$ and $\rho[k]$.

$$Z[k] = (1-\alpha)(Q[k]) + (\alpha)(\rho[k]) \quad (10)$$

As can be deduced from Table 2, the weighted depth density of layers $(Z[k])$ improves the feature pattern for better differentiation for 13 out of 20 features, is the same for 5 features (I through V on standing-walking), and is worse for 2 features (VIII on side-view sitting and X on back-view sitting). Thus, $Z[k]$ is generally very useful, though the similar and worse outcomes would require a further multiview fusion process to distinguish the pattern.

2.2.3. Proportion Value. Proportion value (P_v) is a penalty parameter of the model to indicate roughly the proportion of object vertical actions distinguished from horizontal actions. It is the ratio of the width W_h and the height H_h of the object in each view (see the following equation):

$$P_v = \frac{W_h}{H_h} \quad (11)$$

Table 1 mentioned earlier shows the depth profile of action and its features in each view. In general, the feature patterns of each action are mostly similar, though they do exhibit some differences depending on the viewpoints. It should be noted here that the camera(s) are positioned at 2 m above the floor, pointing down at an angle of 30° from the horizontal line.

Four actions in Table 1, standing/walking, sitting, stooping, and lying, shall be elaborated here. For standing/walking, the features are invariant to viewpoints for both density $(\rho[k])$ and inverse depth $(D_i[k])$. The $D[k]$ values slope equally for every viewpoint due to the position of the camera(s). For sitting, the features vary for $\rho[k]$ and $D[k]$ according to their viewpoints. However, the patterns of sitting for front and slant views are rather similar to standing/walking. $D[k]$ from some viewpoints indicates the hidden volume of the thigh. For stooping, the $\rho[k]$ patterns are quite the same from most viewpoints, except for the front view and back view, due to occlusion of the upper body. However, $D[k]$ reveals clearly the volume in stooping. For lying, the $\rho[k]$ patterns vary depending on the viewpoints and cannot be distinguished using layer-based features. In this particular case, the proportion value (P_v) is introduced to help identify the action.

2.3. Layer-Based Feature Fusion. In this section, we emphasize the fusion of features from various views. The pattern of features can vary or self-overlap with respect to the viewpoint. In a single view, this problem leads to similar and unclear features between actions. Accordingly, we have introduced a method for the fusion of features from multiviews to improve action recognition. From each viewpoint, three features of action are extracted: (1) density of layer $(\rho[k])$, (2) weighted depth density of layers $(Z[k])$, and (3) proportion value (P_v). We have established two fused features as the combination of width, area, and volume of the body structure from every view with respect to layers. These two fused features are the mass of dimension $(\omega[k])$ and the weighted mass of dimension $(\overline{\omega}[k])$.

(a) The mass of dimension feature $(\omega[k])$ is computed from the product of density of layers $(\rho[k])$ in every

Table 1: Sampled human profile depth in various views and actions.

Action/view	I_{mo}	ρ [k]	D [k]	D_i[k]	Action/view	I_{mo}	ρ [k]	D [k]	D_i[k]
Standing-walking/front					Stooping/front				
Standing-walking/slant					Stooping/slant				
Standing-walking/side					Stooping/side				
Standing-walking/back slant					Stooping/back slant				
Standing-walking/back					Stooping/back				
Sitting/front					Lying/front				
Sitting/slant					Lying/slant				

Table 1: Continued.

Action/view	I_{mo}	ρ [k]	D [k]	D_i[k]	Action/view	I_{mo}	ρ [k]	D [k]	D_i[k]
Sitting/side					Lying/side				
Sitting/back slant					Lying/back slant				
Sitting/back					Lying/back				

Note: (I) $\rho[k]$ is density of layer. (II) D $[k]$ is real-range depth of layer. (III) $D_i[k]$ is inverse depth of layer.

Table 2: Sampled features in single view and fused features in multiview.

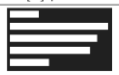
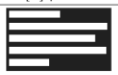
Action/view	Camera 1		Camera 2		$\overline{\omega}[k]$
	I_{mo}	$Z[k] \mid \alpha = 0.9$	I_{mo}	$Z[k] \mid \alpha = 0.9$	
(I) Standing-walking/front					
(II) Standing-walking/slant					
(III) Standing-walking/side					
(IV) Standing-walking/back slant					
(V) Standing-walking/back					
(VI) Sitting/front					
(VII) Sitting/slant					
(VIII) Sitting/side					
(IX) Sitting/back slant					
(X) Sitting/back					
(XI) Stooping/front					
(XII) Stooping/slant					
(XIII) Stooping/side					
(XIV) Stooping/back slant					
(XV) Stooping/back					

Table 2: Continued.

Action/view	Camera 1		Camera 2		$\overline{\omega}[k]$
	I_{mo}	$Z[k] \mid \alpha = 0.9$	I_{mo}	$Z[k] \mid \alpha = 0.9$	
(XVI) Lying/front					
(XVII) Lying/slant					
(XVIII) Lying/side					
(XIX) Lying/back slant					
(XX) Lying/back					

Note: (I) $Z[k]$ is weighted depth density of layers which uses α for adjustment between $Q[k]$ and $\rho[k]$. (II) $\overline{\omega}[k]$, fused features from multiview, called the weighted mass of dimension feature.

view, from $v = 1$ (the first view) to $v = d$ (the last view), as shown in the following equation:

$$\omega[k] = \prod_{v=1}^{d} \left(\rho_{(v=i)}[k]\right) \tag{12}$$

(b) The weighted mass of dimension feature ($\overline{\omega}[k]$) in (13) is defined as the product of the weighted depth density of layers ($Z[k]$) from every view.

$$\overline{\omega}[k] = \prod_{v=1}^{d} \left(Z_{(v=i)}[k]\right) \tag{13}$$

In addition, the maximum proportion value is selected from all views (see (14)) which is used to make feature vector.

$$P_{\mathrm{m}} = \operatorname{argmax}\left(P_{\mathrm{v}(v=1)}, P_{\mathrm{v}(v=2),}, P_{\mathrm{v}(v=3),}, \cdots, P_{\mathrm{v}(v=d)}\right) \tag{14}$$

Two feature vectors, a nondepth feature vector and a depth feature vector, are now defined for use in the classification process.

The nondepth feature vector is formed by concatenating the mass of dimension features ($\omega[k]$) and the maximum proportion values as follows.

$$\{\omega[-N], \omega[-N+1], \omega[-N+2], \ldots, \omega[0], \omega[1], \omega[2], \ldots, \omega[N], P_m\} \tag{15}$$

The depth feature vector is formed by the weighted mass of dimension features ($\overline{\omega}[k]$) concatenated with the maximum proportion values (P_{m}) as follows:

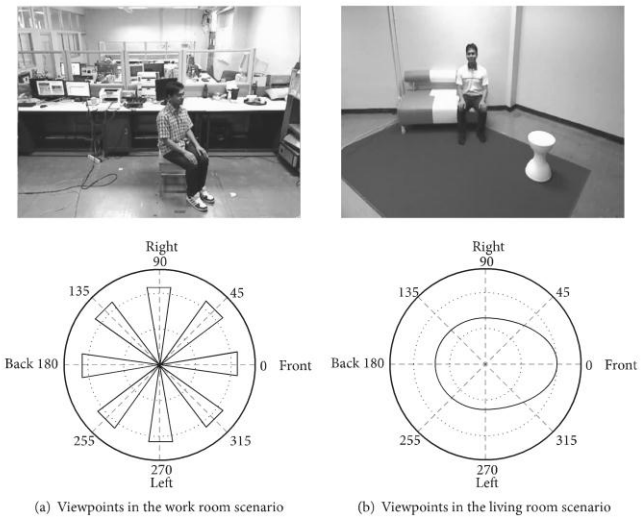
$$\{\overline{\omega}[-N], \overline{\omega}[-N+1], \overline{\omega}[-N+2], \ldots, \overline{\omega}[0], \overline{\omega}[1], \overline{\omega}[2], \ldots, \overline{\omega}[N], P_m\} \tag{16}$$

Table 2 shows the weighted mass of dimension feature ($\overline{\omega}[k]$) fused from the weighted depth density of layers ($Z[k]$) from two cameras. The fused patterns for standing-walking in each view are very similar. For sitting, generally the patterns are more or less similar except in the back view due to the lack of leg images. The classifier, however, can differentiate posture using the thigh part, though there are some variations in the upper body. For stooping, the fused patterns are consistent in all views: heap in the upper part but slightly different in varying degrees of curvature. For lying, all fused feature patterns are different. The depth profiles, particularly in front views, affect the feature adjustment. However, the classifier can still distinguish appearances, because generally patterns of features in the upper layers are shorter than those in the lower layers.

3. Experimental Results

Experiments to test the performance of our method were performed on three datasets: the PSU (Prince of Songkla University) dataset, the NW-UCLA (Northwestern-University of California at Los Angeles) dataset, and the i3DPost dataset. We use the PSU dataset to estimate the optimal parameters in our model, such as the number of layers and the adjustable parameter α. The tests are performed on single and multiviews, angles between cameras, and classification methods. Subsequently, our method is tested using the NW-UCLA dataset and the i3DPost dataset, which is set up from different viewpoints and angles between cameras to evaluate the robustness of our model.

3.1. Experiments on the PSU Dataset. The PSU dataset [76] contains 328 video clips of human profiles with four basic actions recorded in two views using RGBD cameras (Kinect



(a) Viewpoints in the work room scenario (b) Viewpoints in the living room scenario

FIGURE 4: Example of two multiview scenarios of profile-based action for the PSU dataset.

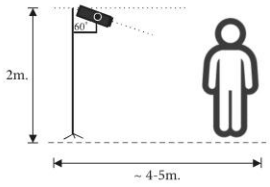


FIGURE 5: General camera installation and setup.

ver.1). The videos were simultaneously captured and synchronized between views. The profile-based actions consisted of standing/walking, sitting, stooping, and lying. Two scenarios, one in a work room for training and another in a living room for testing, were performed. Figure 4 shows an example of each scenario, together with the viewpoints covered.

Two Kinect cameras were set overhead at 60° to the vertical line, each at the end of a 2 m pole. RGB and depth information from multiviews was taken from the stationary cameras with varying viewpoints to observe the areas of interest. The operational range was about 3-5.5 m from the cameras to accommodate a full body image, as illustrated in Figure 5. The RGB resolution for the video dataset was 640×480, while depths at 8 and 24 bits were also of the same resolution. Each sequence was performed by 3-5 actors, having no less than 40 frames of background at the beginning to allow motion detection using any chosen background subtraction technique. The frame rate was about 8-12 fps.

(i) Scenario in the Work Room (Training Set). As illustrated in Figure 6(a), the two cameras' views are perpendicular to each other. There are five angles of object orientation: front (0°), slant (45°), side (90°), rear-slant (135°), and rear (180°), as shown in Figure 6(b). A total of 8,700 frames were obtained in this scenario for training.

(ii) Scenario in the Living Room (Testing Set). This scenario, illustrated in Figure 7, involves one moving Kinect camera at four angles: 30°, 45°, 60°, and 90°, while another Kinect camera remains stationary. Actions are performed freely in various directions and positions within the area of interest. A total of 10,720 frames of actions were tested.

3.1.1. Evaluation of the Number of Layers. We determine the appropriate number of layers by testing our model with different numbers of layers (L) using the PSU dataset. The numbers of layers for testing are 3, 5, 7, 9, 11, 13, 15, 17, and 19. The alpha value (α) is instinctively fixed at 0.7 to find the optimal value of the number of layers. Two classification methods are used for training and testing: an artificial neural network (ANN) with a back-propagation algorithm over 20 nodes of hidden layers and a support vector machine (SVM) using a radial basis function kernel along with C-SVC. The results using ANN and SVM are shown in Figures 8 and 9, respectively.

Figure 8 shows that the 3-layer size achieves the highest average precision of 94.88% using the ANN and achieves 92.11% using the SVM in Figure 9. Because the ANN performs better, in the tests that follow, the evaluation of our

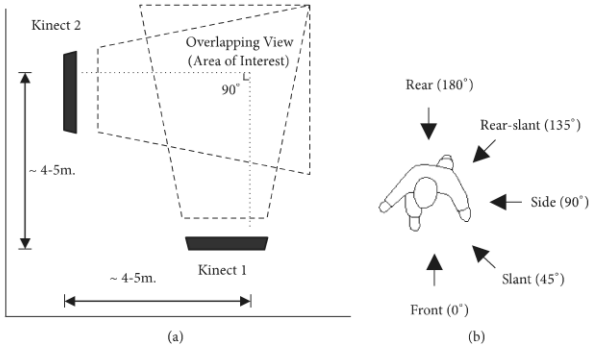


Figure 6: Scenario in the work room. (a) Installation and scenario setup from top view. (b) Orientation angle to object.

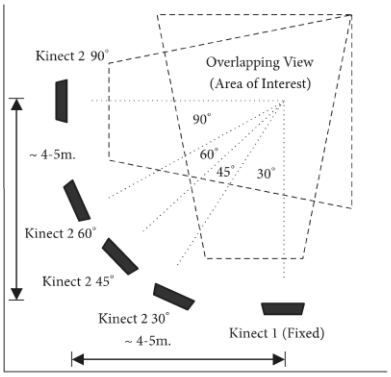


Figure 7: Installation and scenario of the living room setup (four positions of Kinect camera 2, at 30°, 45°, 60°, and 90°, and stationary Kinect camera 1).

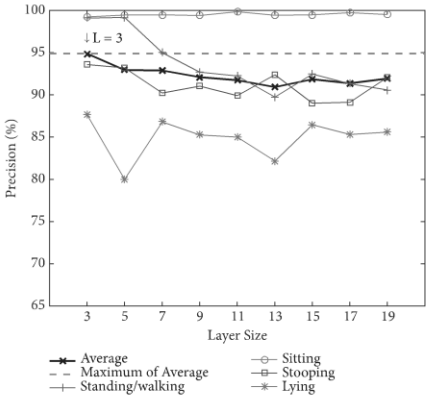


Figure 8: Precision by layer size for the four postures using an artificial neural network (ANN) on the PSU dataset.

model will be based on this layer size together with the use of the ANN classifier.

3.1.2. Evaluation of Adjustable Parameter α. For the weighted mass of dimension feature, ($\overline{\omega}[k]$), the adjustment parameter alpha (α)—the value for the weight between the inverse depth density of layers (Q[k]) and the density of layer ($\rho[k]$)—is employed. The optimal α value is determined to show the improved feature that uses inverse depth to reveal hidden volume in some parts and the normal volume. The experiment is carried out by varying alpha from 0 to 1 at 0.1 intervals.

Figure 10 shows the precision of action recognition using a 3-layer size and the ANN classifier versus the alpha values. In general, except for the sitting action, one may note that as the portion of inverse depth density of layers (Q[k]) is augmented, precision increases. The highest average precision is 95.32% at $\alpha = 0.9$, meaning that 90% of the inverse depth density of layers (Q[k]) and 10% of the density of layers ($\rho[k]$) are an optimal proportion. When α is above 0.9, all precision values drop. The trend for sitting action is remarkably different from others in that precision always hovers near the maximum and gradually but slightly decreases when α increases.

Figure 11 illustrates the multiview confusion matrix of action precisions when L = 3 and $\alpha = 0.9$, using the PSU dataset. We found that standing/walking action had the highest precision (99.31%), while lying only reached 90.65%. The classification error of the lying action depends mostly on its characteristic that the principal axis of the body is

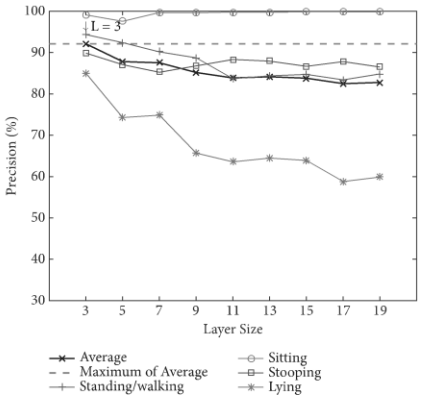


Figure 9: Precision by layer size for the four postures using support vector machine (SVM) on the PSU dataset.

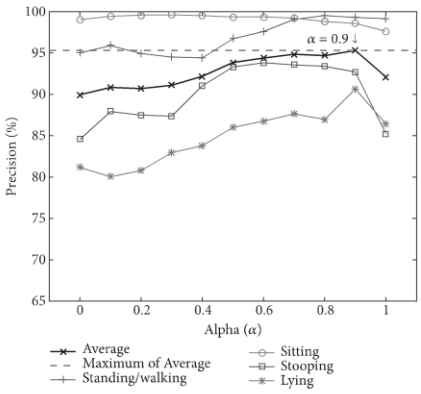


Figure 10: Precision versus α (the adjustment parameter for weighting between $Q[k]$ and $\rho[k]$ when L = 3, from the PSU dataset).

aligned horizontally, which works against the feature model. In general, the missed classifications of standing/walking, stooping, and lying actions were mostly confused with the sitting action, accounting for 0.69%, 4.82%, and 8.83% of classifications, respectively. Nevertheless, the precision of sitting is relatively high at 98.59%.

3.1.3. Comparison of Single View/Multiview. For the sake of comparison, we also evaluate tests in single-view recognition for the PSU dataset for L = 3 and α = 0.9 in the living room scene, similar to that used to train the classification model for the work room. Figure 12 shows the results from

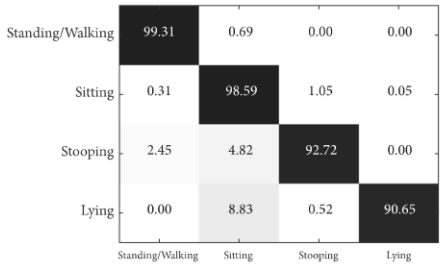


Figure 11: Confusion matrix for multiview recognition in the PSU dataset when L = 3 and α = 0.9.

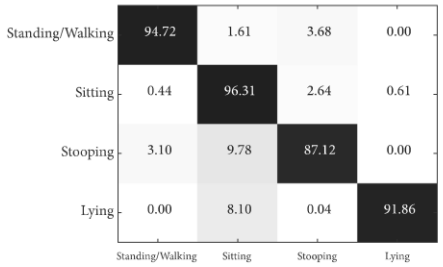


Figure 12: Confusion matrix of single-view Kinect 1 recognition (stationary camera) for the PSU dataset when L = 3 and α = 0.9.

the single-view Kinect 1 (stationary camera), while Figure 13 shows those from the single-view Kinect 2 (moving camera).

Results show that the single-view Kinect 1, which is stationary, performs slightly better than the single-view Kinect 2, which is moving (average precision of 92.50% compared to 90.63%). The stationary camera gives the best results for sitting action, while for the moving camera, the result is best for the standing/walking action. It is worth noting that the stationary camera yields a remarkably better result for lying action than the moving one.

Figure 14 shows the precision of each of the four postures, together with that of the average, for the multiview and the two single views. On average, the result is best accomplished with the use of the multiview and is better for all postures other than the lying action. In this regard, single-view 1 yields a slightly better result, most probably due to its stationary viewpoint toward the sofa, which is perpendicular to its line of sight.

3.1.4. Comparison of Angle between Cameras. As depicted earlier in Figure 7, the single-view Kinect 1 camera is stationary, while the single-view Kinect 2 camera is movable, adjusted to capture viewpoints at 30°, 45°, 60°, and 90° to the stationary camera. We test our model on these angles to assess

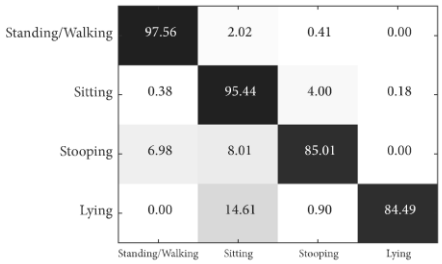


Figure 13: Confusion matrix of single-view Kinect 2 recognition (moving camera) for the PSU dataset when L = 3 and $\alpha = 0.9$.

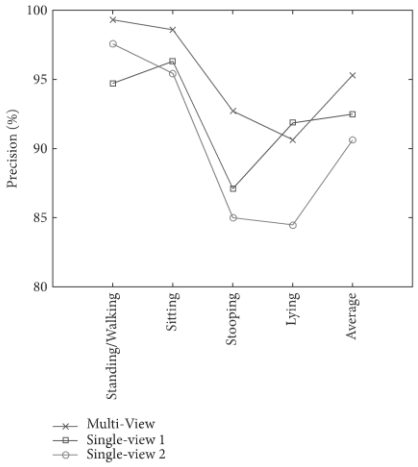


Figure 14: Comparison of precision from multiview, single-view 1, and single-view 2 for the PSU dataset when L = 3 and $\alpha = 0.9$.

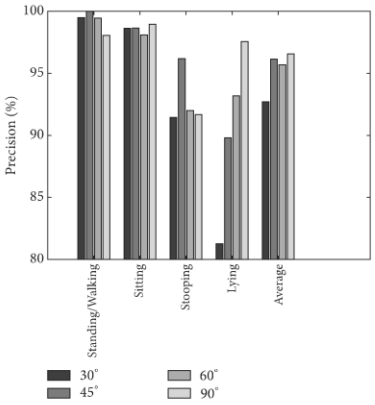


Figure 15: Precision comparison graph on different angles in each action.

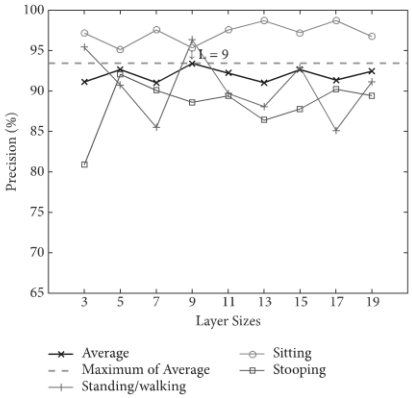


Figure 16: Precision by layer size for the PSU dataset when using the NW-UCLA-trained model.

the robustness of the model on the four postures. Results are shown in Figure 15.

In Figure 15, the lowest average precision result occurs at 30°—the smallest angle configuration between the two cameras. This is most probably because the angle is narrow and thus not much additional information is gathered. For all other angles, the results are closely clustered. In general, standing/walking and sitting results are quite consistent at all angles, while lying and stooping are more affected by the change.

3.1.5. Evaluation with NW-UCLA Trained Model. In addition, we tested the PSU dataset in living room scenes using the model trained on the NW-UCLA dataset [63]. Figure 16 illustrates that the 9-layer size achieves the highest average precision at 93.44%. Sitting gives the best results and the highest precision, up to 98.74% when L = 17. However, sitting at low layers also gives good results, for example, L = 3 at 97.18%, while the highest precision for standing/walking is 95.40%, and the lowest precision is 85.16%. The lowest precision for stooping is 92.08% when L = 5.

Figure 17 shows the results when L = 9, which illustrates that standing/walking gives the highest precision at 96.32%, while bending gives the lowest precision at 88.63%.

In addition, we also compare precision of different angles between cameras, as shown in Figure 18. The result shows that the highest precision on average is 94.74% at 45°, and the lowest is 89.69% at 90°.

Table 3: Time consumption testing on different numbers of layers and different classifiers.

Classifier	Time consumption (ms)/frame									Frame rate (fps)		
	L=3			L=11			L=19			L=3	L=11	L=19
	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Avg	Avg	Avg
ANN	13.95	17.77	15.08	14.12	20.58	15.18	14.43	20.02	15.85	66.31	65.88	63.09
SVM	14.01	17.40	15.20	14.32	18.97	15.46	14.34	19.64	15.71	65.79	64.68	63.65

Note: L stands for the number of layers. ANN, artificial neural network; SVM, support vector machine.

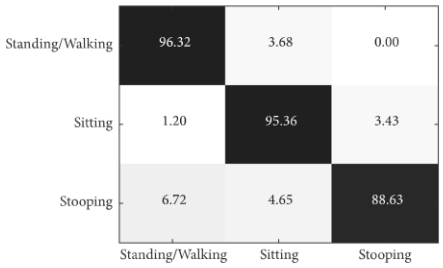


Figure 17: Confusion matrix of 2 views for the PSU dataset (using the NW-UCLA-trained model) when L = 9.

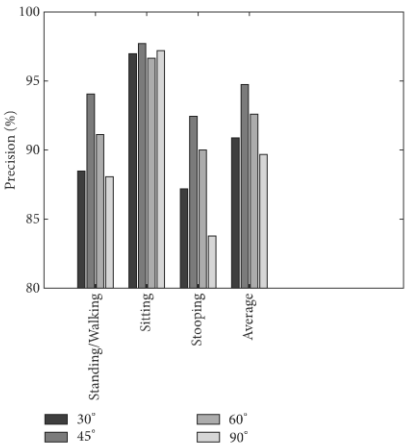


Figure 18: Precision comparison graph of different angles in each action when using the NW-UCLA-trained model.

In general, the precision of all actions is highest at 45° and decreases as the angle becomes larger. However, the results obtained using the PSU-trained model show a different trend, where a larger angle provides better results.

3.1.6. Evaluation of Time Consumption. Time consumption evaluation, excluding interface and video showing time, is conducted using the OpenMP wall clock. Our system is tested on a normal PC (Intel® Core™ i5 4590 at 3.30 GHz with 8 GB DDR3). We use the OpenCV library for computer vision, the OpenMP library for parallel processing, and CLNUI to capture images from the RGBD cameras. The number of layers and the classifier are tested using 10,720 action frames from the living room scene.

On average, the time consumption is found to be approximately 15 ms per frame or a frame rate of around 63 fps. As detailed in Table 3, the number of layers and the type of classifier affect the performance only slightly. In addition, we compare serial processing with parallel processing, which divides a single process into threads. The latter is found to be 1.5507 times faster than the former. It was noted that thread initialization and synchronization consume a portion of the computation time.

3.2. Experiment on the NW-UCLA Dataset. The NW-UCLA dataset [63] is used to benchmark our method. This dataset is similar to our work for the multiview action 3D PSU dataset taken at different viewpoints to capture the RGB and depth images. The NW-UCLA dataset covers nearly ten actions including *stand up*, *walk around*, *sit down*, and *bend to pick up item*, but it lacks a *lie down* action. Actions in this dataset are marked in colors. To test our method, the motion detection step for segmenting movement is replaced by a specific color segmentation to obtain the human structure. The rest of the procedure stays the same.

Only actions of interest are selected for the test—its transition actions are excluded. For example, standing/walking frames are extracted from stand up and walk around; sitting frames are selected from sit down and stand up; and stooping is extracted from *pick up with one hand* and *pick up with both hands*.

Our method employs the model learned using the PSU dataset in the work room scenario to test the NW-UCLA dataset. All parameters in the test are the same except that alpha is set to zero due to variations in depth values. The experiment is performed for various numbers of layers from L = 3 to L = 19. Test results on the NW-UCLA dataset are shown in Figure 19.

From Figure 19, the maximum average precision (86.40%) of the NW-UCLA dataset is obtained at layer L =11, in contrast to the PSU dataset at L = 3. Performance for stooping is generally better than other actions and peaks at 95.60%. As detailed in Figure 20, standing/walking gives the lowest precision at 76.8%. The principal cause of low

Table 4: Comparison between NW-UCLA and our recognition systems on the NW-UCLA dataset.

NW-UCLA recognition action [63]	Precision (%)	Our recognition action	Precision (%)
Walk around	77.60	Walking/standing	76.80
Sit down	68.10	Sitting	86.80
Bend to pick item (1 hand)	74.50	Stooping	95.60
Average	73.40	Average	86.40

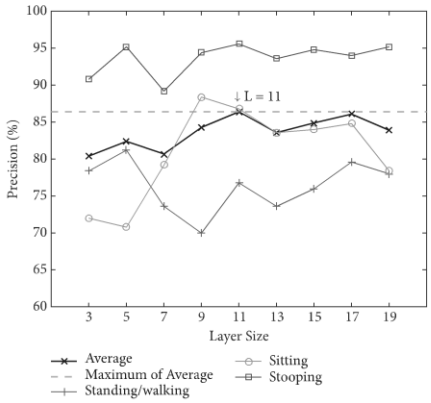


Figure 19: Precision of action by layer size on NW-UCLA dataset.

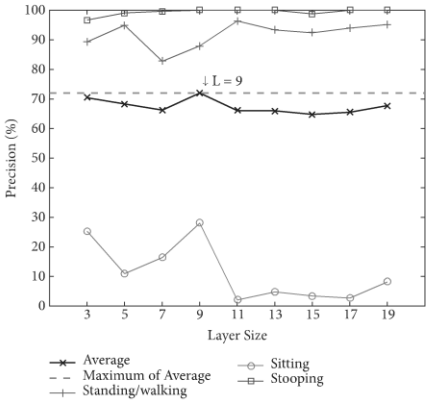


Figure 21: Precision of i3DPost dataset by layer size when using the PSU-trained model.

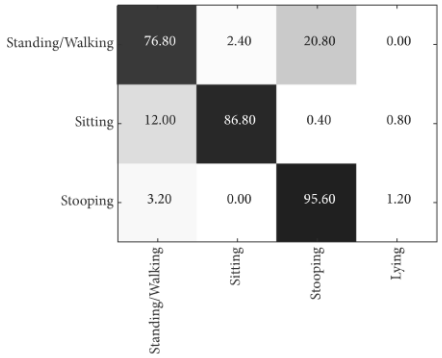


Figure 20: Confusion matrix of test results on the NW-UCLA 3D dataset when L = 11.

precision is that the angle of the camera and its captured range are very different from the PSU dataset. Compared with the method proposed by NW-UCLA, our method performs better by up to 13 percentage points, from an average of 73.40% to 86.40%, as shown in Table 4. However, to be fair, many more activities are considered by the NW-UCLA than ours which focuses only on four basic actions and hence its disadvantage by comparison.

3.3. Experiment on the i3DPost Dataset. The i3DPost dataset [77] is an RGB multiview dataset that contains 13 activities. The dataset was captured by 8 cameras from different viewpoints with 45° between cameras, performed by eight persons for two sets at different positions. The background images allow the same segmentation procedure to build the *nondepth feature vector* for recognizing profile-based action.

The testing sets extract only target actions from temporal activities, including standing/walking, sitting, and stooping from *sit-standup*, *walk*, *walk-sit*, and *bend*.

3.3.1. Evaluation of i3DPost Dataset with PSU-Trained Model. Firstly, i3DPost is tested by the PSU-trained model from 2 views at 90° between cameras on different layer sizes.

Figure 21 shows the testing results of the i3DPost dataset using the PSU-trained model. The failed prediction for sitting shows a precision of only 28.08% at L = 9. By observation, the mistake is generally caused by the action of sitting that looks like a squat in the air, which is predicted as standing. In the PSU dataset, sitting is done on a bench/chair. On the other hand, standing/walking and stooping are performed with good results of about 96.40% and 100% at L = 11, respectively.

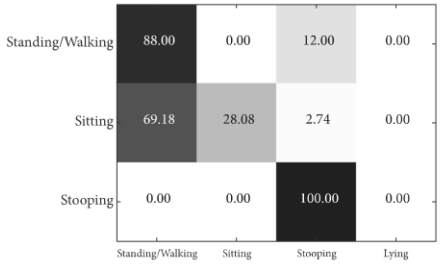


Figure 22: Confusion matrix of 2 views for i3DPost dataset using PSU-trained model when L = 9.

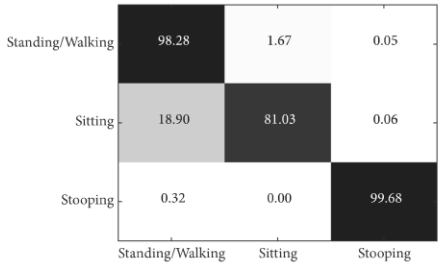


Figure 24: Confusion matrix of test result on the 2-view i3DPost dataset when L = 17.

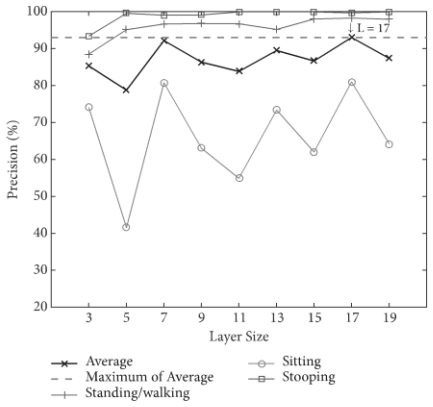


Figure 23: Precision of i3DPost dataset with new trained model by layer size.

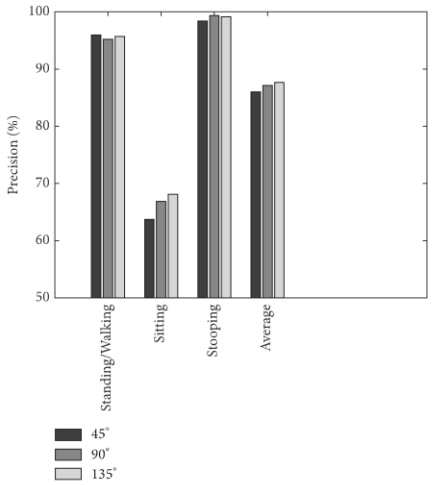


Figure 25: Precision comparison graph for different angles in the i3DPost.

Figure 22 shows the multiview confusion matrix when L = 9. Sitting is most often confused with standing/walking (69.18% of cases), and standing/walking is confused with stooping (12.00% of cases).

3.3.2. Training and Evaluation Using i3DPost. From the last section, i3DPost evaluation using the PSU-trained model resulted in missed classification for the sitting action. Accordingly, we experimented with our model by training and evaluating using only the i3DPost; the first dataset is used for testing and the second for training. The initial testing is performed in 2 views.

Figures 23 and 24 show the results of each layer size from 2 views. The 17 layers achieve the highest precision at 93.00% on average (98.28%, 81.03%, and 99.68% for standing/walking, sitting, and stooping, resp.). In general, standing/walking and stooping achieve good precision of above 90%, except at L = 3. However, the best precision for sitting is only 81.03%, where most wrong classifications are defined as standing/walking (18.90% of cases), and the lowest precision is at 41.59% when L = 5. We noticed that the squat still affects performance.

In 2-view testing, we couple the views for different angles between cameras, such as 45°, 90°, and 135°. Figure 25 shows that, at 135°, the performance on average is highest, and the lowest performance is at 45°. In general, a smaller angle gives lower precision; however, for sitting, a narrow angle may reduce precision dramatically.

In addition, we perform the multiview experiments for various numbers of views from one to six in order to evaluate our multiview model, as shown in Figure 26.

Figure 26 shows the precision according to the number of views. The graph reports the maximum precision versus the different number of views from one to six views, which are 89.03%, 93.00%, 91.33%, 92.30%, 92.56%, and 91.03% at L = 7,

Table 5: Comparison between our recognition systems and [59] on the i3DPost dataset.

Recognition of similar approach [59]	Precision (%)	Our recognition action	Precision (%)
Walk	95.00	Walking/standing	98.28
Sit	87.00	Sitting	81.03
Bend	100.0	Stooping	99.68
Average	94.00	Average	93.00

Table 6: Precision results of our study and others emphasizing profile-based action recognition using different methods and datasets.

Method	Precision rate of action (%)					
	Standing	*Walking*	*Sitting*	*Stooping*	*Lying*	*Average*
P. Chawalitsittikul et al. [70]	98.00	98.00	93.00	94.10	98.00	96.22
N. Noorit et al. [71]	99.41	80.65	89.26	94.35	100.0	92.73
M. Ahmad et al. [72]	-	89.00	85.00	100.0	91.00	91.25
C. H. Chuang et al. [73]	-	92.40	97.60	95.40	-	95.80
G. I. Parisi et al. [74]	96.67	90.00	83.33	-	86.67	89.17
N. Sawant et al. [75]	91.85	96.14	85.03	-	-	91.01
Our method on PSU dataset	99.31	99.31	98.59	92.72	90.65	95.32

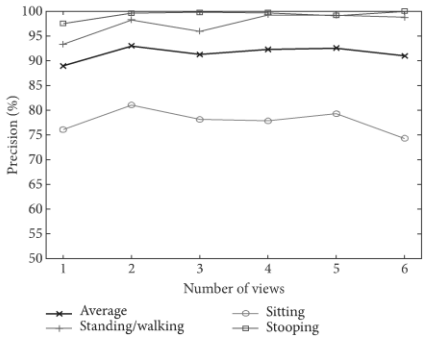


Figure 26: Precision in i3DPost dataset with new trained model by the number of views.

L = 17, L = 17, L = 7, and L = 13, respectively. We noticed that the highest precision is for 2 views. In general, the performance increases when the number of views increases, except for sitting. Moreover, the number of layers that give maximum precision is reduced as the number of views increases. In conclusion, only two or three views from different angles are necessary for obtaining the best performance.

We compared our method with a similar approach [59], based on a posture prototype map and results voting function with a Bayesian framework for multiview fusion. Table 5 shows the comparison results. The highest precisions of our method and the comparison approach are 99.68% and 100% for stooping and bend, respectively. Likewise, the lowest precisions are 81.03% and 87.00% for the same actions. However, for walking/standing, our approach obtains better results. On average, the comparison approach performs slightly better than our method.

4. Comparison with Other Studies

Our study has now been presented with other visual profile-based action recognition studies and also compared with other general action recognition studies.

4.1. Precision Results of Different Studies. Table 6 shows precision results of various methods emphasizing profile-based action recognition. Note that the methods employed different algorithms and datasets; thus results are presented only for the sake of studying their trends with respect to actions. Our method is tested on the PSU dataset. Our precision is highest on walking and sitting actions, quite good on the standing action, quite acceptable on the lying action, but poor on the stooping action. It is not possible to compare on precision due to lack of information on the performance of other methods. However, we can note that each method performs better than others on different actions; for example, [74] is good for standing, and [73] is a better fit for sitting.

4.2. Comparison with Other Action Recognition Studies. Table 7 compares the advantages and disadvantages of some previous studies concerning action recognition acquired according to the criteria of their specific applications. In the inertial sensor-based approach, sensors/devices are attached to the body and hence monitoring is available everywhere, for the inconvenience of carrying them around. This approach gives high privacy but is highly complex. In an RGB vision-based approach, sensors are not attached to the body, and hence it is less cumbersome. Though not so complex, the main drawback of this method is the lack of privacy. In this regard, depth-based views may provide more privacy. Although rather similar in comparison in the table, the multiview approach can cope with some limitations, such

Table 7: Comparison of some action recognition approaches based on ease of use, perspective robustness, and real-world application conditions.

Approach	Flexible calibration	View scalability	Perspective robustness	Real-world Application Condition		Privacy
				Complexity	Cumbersomeness from sensor/device attachment	
Inertial sensor-based [1–12]	N/A	N/A	High	Highly complex	Cumbersome	High
RGB vision-based [13, 17, 21–47, 71–73, 75]	Low-moderate	Low-moderate	Low-moderate	Simple-moderate	Not cumbersome	Lacking
Depth-based Single-view [14–16, 18–20, 70, 74]	Low-moderate	Low-moderate	Low-moderate	Depends	Not cumbersome	Moderate-high
RGB/depth-based multiview [48–68]	Low-moderate	Low-moderate	Low-moderate	Depends	Not cumbersome	Depends
Our work, depth-based multiview	**Moderate-high**	**Moderate-high**	**High**	**Simple**	**Not cumbersome**	**High**

as vision coverage and continuity or obstruction, which are common in the single-view approach, as described in some detail in Introduction.

(i) As for our proposed work, it can be seen from the table that the approach is in general quite good in comparison; in addition to being simple, it offers a high degree of privacy and other characteristics such as flexibility, scalability, and robustness are at similar levels, if not better, and no calibration is needed, which is similar to most of the other approaches.

However, the two cameras for our multiview approach still have to be installed following certain specifications, such as the fact that the angle between cameras should be more than 30°.

5. Conclusion/Summary and Further Work

In this paper, we explore both the inertial and visual approaches to camera surveillance in a confined space such as a healthcare home. The less cumbersome vision-based approach is studied in further detail for single-view and multiview RGB depth-based and for non-RGB depth-based approaches. We decided on a multiview, non-RGB depth-based approach for privacy and have proposed a layer fusion model, which is representation-based model that allows fusion of features and information by segmenting parts of the object into vertical layers.

We trained and tested our model on the PSU dataset with four postures (standing/walking, sitting, stooping, and lying) and have evaluated the outcomes using the NW-UCLA and i3DPost datasets on available postures that could be extracted. Results show that our model achieved an average precision of 95.32% on the PSU dataset—on a par with many other achievements, if not better, 93.00% on the i3DPost dataset, and 86.40% on the NW-UCLA dataset. In addition to flexibility of installation, scalability, and noncalibration of features, one advantage over most other approaches is that our approach is simple, while it contributes good recognition on various viewpoints with a high speed of 63 frames per second, suitable for application in a real-world setting.

In further research, a spatial-temporal feature is interesting and should be investigated for more complex action recognition, such as waving and kicking. Moreover, reconstruction of 3D structural and bag-of-visual-word models is also of interest.

Data Availability

The PSU multiview profile-based action dataset is available at [74], which is authorized only for noncommercial or educational purposes. The additional datasets to support this study are cited at relevant places within the text as references [63] and [75].

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was partially supported by the 2015 Graduate School Thesis Grant, Prince of Songkla University (PSU), and Thailand Center of Excellence for Life Sciences (TCELS), and the authors wish to express their sincere appreciation. The first author is also grateful for the scholarship funding granted for him from Rajamangala University of Technology Srivijaya to undertake his postgraduate study, and he would like to express gratitude to Assistant Professor Dr. Nikom Suvonvorn for his guidance and advices. Thanks are extended to all members of the Machine Vision Laboratory, PSU Department of Computer Engineering, for the sharing of their time in the making of the PSU profile-based action dataset. Last but not least, specials thanks are due to Mr. Wiwat Sutiwipakorn, a former lecturer at the PSU Faculty of Engineering, regarding the use of language in the manuscript.

References

[1] Z. He and X. Bai, "A wearable wireless body area network for human activity recognition," in *Proceedings of the 6th International Conference on Ubiquitous and Future Networks, ICUFN 2014*, pp. 115–119, China, July 2014.

[2] C. Zhu and W. Sheng, "Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 41, no. 3, pp. 569–573, 2011.

[3] J.-S. Sheu, G.-S. Huang, W.-C. Jheng, and C.-H. Hsiao, "Design and implementation of a three-dimensional pedometer accumulating walking or jogging motions," in *Proceedings of the 2nd International Symposium on Computer, Consumer and Control, IS3C 2014*, pp. 828–831, Taiwan, June 2014.

[4] R. Samiei-Zonouz, H. Memarzadeh-Tehran, and R. Rahmani, "Smartphone-centric human posture monitoring system," in *Proceedings of the 2014 IEEE Canada International Humanitarian Technology Conference, IHTC 2014*, pp. 1–4, Canada, June 2014.

[5] X. Yin, W. Shen, J. Samarabandu, and X. Wang, "Human activity detection based on multiple smart phone sensors and machine learning algorithms," in *Proceedings of the 19th IEEE International Conference on Computer Supported Cooperative Work in Design, CSCWD 2015*, pp. 582–587, Italy, May 2015.

[6] C. A. Siebra, B. A. Sa, T. B. Gouveia, F. Q. Silva, and A. L. Santos, "A neural network based application for remote monitoring of human behaviour," in *Proceedings of the 2015 International Conference on Computer Vision and Image Analysis Applications (ICCVIA)*, pp. 1–6, Sousse, Tunisia, January 2015.

[7] C. Pham, "MobiRAR: real-time human activity recognition using mobile devices," in *Proceedings of the 7th IEEE International Conference on Knowledge and Systems Engineering, KSE 2015*, pp. 144–149, Vietnam, October 2015.

[8] G. M. Weiss, J. L. Timko, C. M. Gallagher, K. Yoneda, and A. J. Schreiber, "Smartwatch-based activity recognition: A machine learning approach," in *Proceedings of the 3rd IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2016*, pp. 426–429, USA, February 2016.

[9] Y. Wang and X.-Y. Bai, "Research of fall detection and alarm applications for the elderly," in *Proceedings of the 2013 International Conference on Mechatronic Sciences, Electric Engineering*

and Computer, MEC 2013, pp. 615–619, China, December 2013.

[10] Y. Ge and B. Xu, "Detecting falls using accelerometers by adaptive thresholds in mobile devices," *Journal of Computers in Academy Publisher*, vol. 9, no. 7, pp. 1553–1559, 2014.

[11] W. Liu, Y. Luo, J. Yan, C. Tao, and L. Ma, "Falling monitoring system based on multi axial accelerometer," in *Proceedings of the 2014 11th World Congress on Intelligent Control and Automation (WCICA)*, pp. 7–12, Shenyang, China, June 2014.

[12] J. Yin, Q. Yang, and J. J. Pan, "Sensor-based abnormal human-activity detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 8, pp. 1082–1090, 2008.

[13] B. X. Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 1293–1301, Boston, Mass, USA, June 2015.

[14] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *Proceedings of the 22nd International Conference on Pattern Recognition, ICPR 2014*, pp. 4513–4518, Sweden, August 2014.

[15] E. Ohn-Bar and M. M. Trivedi, "Joint angles similarities and HOG2 for action recognition," in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2013*, pp. 465–470, USA, June 2013.

[16] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 588–595, Columbus, Ohio, USA, June 2014.

[17] V. Parameswaran and R. Chellappa, "View invariance for human action recognition," *International Journal of Computer Vision*, vol. 66, no. 1, pp. 83–101, 2006.

[18] G. Lu, Y. Zhou, X. Li, and M. Kudo, "Efficient action recognition via local position offset of 3D skeletal body joints," *Multimedia Tools and Applications*, vol. 75, no. 6, pp. 3479–3494, 2016.

[19] A. Tejero-de-Pablos, Y. Nakashima, N. Yokoya, F.-J. Díaz-Pernas, and M. Martínez-Zarzuela, "Flexible human action recognition in depth video sequences using masked joint trajectories," *Eurasip Journal on Image and Video Processing*, vol. 2016, no. 1, pp. 1–12, 2016.

[20] E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante, "A Human activity recognition system using skeleton data from RGBD sensors," *Computational Intelligence and Neuroscience*, vol. 2016, Article ID 4351435, 2016.

[21] P. Peursum, H. H. Bui, S. Venkatesh, and G. West, "Robust recognition and segmentation of human actions using HMMs with missing observations," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 13, pp. 2110–2126, 2005.

[22] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Journal of Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[23] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[24] P. Matikainen, M. Hebert, and R. Sukthankar, "Trajectons: Action recognition through the motion analysis of tracked features," in *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops 2009*, pp. 514–521, Japan, October 2009.

[25] M. Jain, H. Jegou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013*, pp. 2555–2562, Portland, OR, USA, June 2013.

[26] S. Zhu and L. Xia, "Human action recognition based on fusion features extraction of adaptive background subtraction and optical flow model," *Journal of Mathematical Problems in Engineering*, vol. 2015, pp. 1–11, 2015.

[27] D.-M. Tsai, W.-Y. Chiu, and M.-H. Lee, "Optical flow-motion history image (OF-MHI) for action recognition," *Journal of Signal, Image and Video Processing*, vol. 9, no. 8, pp. 1897–1906, 2015.

[28] L. Wang, Y. Qiao, and X. Tang, "MoFAP: a multi-level representation for action recognition," *International Journal of Computer Vision*, vol. 119, no. 3, pp. 254–271, 2016.

[29] W. Kim, J. Lee, M. Kim, D. Oh, and C. Kim, "Human action recognition using ordinal measure of accumulated motion," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, Article ID 219190, 2010.

[30] W. Zhang, Y. Zhang, C. Gao, and J. Zhou, "Action recognition by joint spatial-temporal motion feature," *Journal of Applied Mathematics*, vol. 2013, pp. 1–9, 2013.

[31] H.-B. Tu, L.-M. Xia, and Z.-W. Wang, "The complex action recognition via the correlated topic model," *The Scientific World Journal*, vol. 2014, Article ID 810185, 10 pages, 2014.

[32] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[33] A. Yilmaz and M. Shah, "A differential geometric approach to representing the human actions," *Computer Vision and Image Understanding*, vol. 109, no. 3, pp. 335–351, 2008.

[34] M. Grundmann, F. Meier, and I. Essa, "3D shape context and distance transform for action recognition," in *Proceedings of the 2008 19th International Conference on Pattern Recognition, ICPR 2008*, pp. 1–4, USA, December 2008.

[35] D. Batra, T. Chen, and R. Sukthankar, "Space-time shapelets for action recognition," in *Proceedings of the 2008 IEEE Workshop on Motion and Video Computing, WMVC*, pp. 1–6, USA, January 2008.

[36] S. Sadek, A. Al-Hamadi, G. Krell, and B. Michaelis, "Affine-invariant feature extraction for activity recognition," *International Scholarly Research Notices on Machine Vision*, vol. 2013, pp. 1–7, 2013.

[37] C. Achard, X. Qu, A. Mokhber, and M. Milgram, "A novel approach for recognition of human actions with semi-global features," *Machine Vision and Applications*, vol. 19, no. 1, pp. 27–34, 2008.

[38] L. Díaz-Más, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and R. Medina-Carnicer, "Three-dimensional action recognition using volume integrals," *Pattern Analysis and Applications*, vol. 15, no. 3, pp. 289–298, 2012.

[39] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Spatiotemporal features for action recognition and salient event detection," *Cognitive Computation*, vol. 3, no. 1, pp. 167–184, 2011.

[40] N. Ikizler and P. Duygulu, "Histogram of oriented rectangles: A new pose descriptor for human action recognition," *Image and Vision Computing*, vol. 27, no. 10, pp. 1515–1526, 2009.

[41] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, Anchorage, Alaska, USA, June 2008.

[42] V. Kellokumpu, G. Zhao, and M. Pietikainen, "Human activity recognition using a dynamic texture based method," in *Proceedings of the British Machine Vision Conference 2008*, pp. 885–894, Leeds, England, UK, 2008.

[43] D. Tran, A. Sorokin, and D. A. Forsyth, "Human activity recognition with metric learning," in *Proceedings of the European Conference on Computer Vision (ECCV '08)*, Lecture Notes in Computer Science, pp. 548–561, Springer, 2008.

[44] B. Wang, Y. Liu, W. Wang, W. Xu, and M. Zhang, "Multi-scale locality-constrained spatiotemporal coding for local feature based human action recognition," *The Scientific World Journal*, vol. 2013, Article ID 405645, 11 pages, 2013.

[45] A. Gilbert, J. Illingworth, and R. Bowden, "Action recognition using mined hierarchical compound features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 883–897, 2011.

[46] I. C. Duta, J. R. R. Uijlings, B. Ionescu, K. Aizawa, A. G. Hauptmann, and N. Sebe, "Efficient human action recognition using histograms of motion gradients and VLAD with descriptor shape information," *Multimedia Tools and Applications*, vol. 76, no. 21, pp. 22445–22472, 2017.

[47] M. Ahad, M. Islam, and I. Jahan, "Action recognition based on binary patterns of action-history and histogram of oriented gradient," *Journal on Multimodal User Interfaces*, vol. 10, no. 4, pp. 335–344, 2016.

[48] S. Pehlivan and P. Duygulu, "A new pose-based representation for recognizing actions from multiple cameras," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 140–151, 2011.

[49] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 3209–3216, June 2011.

[50] N. Gkalelis, N. Nikolaidis, and I. Pitas, "View indepedent human movement recognition from multi-view video exploiting a circular invariant posture representation," in *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo, ICME 2009*, pp. 394–397, USA, July 2009.

[51] R. Souvenir and J. Babbs, "Learning the viewpoint manifold for action recognition," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, USA, June 2008.

[52] M. Ahmad and S. W. Lee, "HMM-based human action recognition using multiview image sequences," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*, pp. 263–266, Hong Kong, China, 2006.

[53] A. Yao, J. Gall, and L. Van Gool, "Coupled action recognition and pose estimation from multiple views," *International Journal of Computer Vision*, vol. 100, no. 1, pp. 16–37, 2012.

[54] X. Ji, Z. Ju, C. Wang, and C. Wang, "Multi-view transition HMMs based view-invariant human action recognition method," *Multimedia Tools and Applications*, vol. 75, no. 19, pp. 11847–11864, 2016.

[55] A. K. S. Kushwaha, S. Srivastava, and R. Srivastava, "Multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns," *Journal of Multimedia Systems*, vol. 23, no. 4, pp. 451–467, 2017.

[56] S. Spurlock and R. Souvenir, "Dynamic view selection for multi-camera action recognition," *Machine Vision and Applications*, vol. 27, no. 1, pp. 53–63, 2016.

[57] A. A. Chaaraoui and F. Flórez-Revuelta, "A low-dimensional radial silhouette-based feature for fast human action recognition fusing multiple views," *International Scholarly Research Notices*, vol. 2014, pp. 1–11, 2014.

[58] A. Iosifidis, A. Tefas, and I. Pitas, "View-independent human action recognition based on multi-view action images and discriminant learning," in *Proceeding of the IEEE Image, Video, and Multidimensional Signal Processing Workshop 2013*, pp. 1–4, 2013.

[59] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 412–424, 2012.

[60] R. Kavi, V. Kulathumani, F. Rohit, and V. Kecojevic, "Multiview fusion for activity recognition using deep neural networks," *Journal of Electronic Imaging*, vol. 25, no. 4, Article ID 043010, 2016.

[61] Y. Kong, Z. Ding, J. Li, and Y. Fu, "Deeply learned view-invariant features for cross-view action recognition," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 3028–3037, 2017.

[62] A.-A. Liu, N. Xu, Y.-T. Su, H. Lin, T. Hao, and Z.-X. Yang, "Single/multi-view human action recognition via regularized multi-task learning," *Journal of Neurocomputing*, vol. 151, no. 2, pp. 544–553, 2015.

[63] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning, and recognition," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014*, pp. 2649–2656, USA, June 2014.

[64] P. Huang, A. Hilton, and J. Starck, "Shape similarity for 3D video sequences of people," *International Journal of Computer Vision*, vol. 89, no. 2-3, pp. 362–381, 2010.

[65] A. Veeraraghavan, A. Srivastava, A. K. Roy-Chowdhury, and R. Chellappa, "Rate-invariant recognition of humans and their activities," *IEEE Transactions on Image Processing*, vol. 18, no. 6, pp. 1326–1339, 2009.

[66] I. N. Junejo, E. Dexter, I. Laptev, and P. Pérez, "View-independent action recognition from temporal self-similarities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 172–185, 2011.

[67] A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 347–360, 2012.

[68] H. Rahmani and A. Mian, "3D action recognition from novel viewpoints," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 1506–1515, USA, July 2016.

[69] P. KaewTraKuPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Proceeding of Advanced Video-Based Surveillance Systems Springer*, pp. 135–144, 2001.

[70] P. Chawalitsittikul and N. Suvonvorn, "Profile-based human action recognition using depth information," in *Proceedings of the IASTED International Conference on Advances in Computer Science and Engineering, ACSE 2012*, pp. 376–380, Thailand, April 2012.

[71] N. Noorit, N. Suvonvorn, and M. Karnchanadecha, "Model-based human action recognition," in *Proceedings of the 2nd International Conference on Digital Image Processing*, Singapore, February 2010.

[72] M. Ahmad and S.-W. Lee, "Human action recognition using shape and CLG-motion flow from multi-view image

sequences," *Journal of Pattern Recognition*, vol. 41, no. 7, pp. 2237–2252, 2008.

[73] C.-H. Chuang, J.-W. Hsieh, L.-W. Tsai, and K.-C. Fan, "Human action recognition using star templates and delaunay triangulation," in *Proceedings of the 2008 4th International Conference on Intelligent Information Hiding and Multiedia Signal Processing, IIH-MSP 2008*, pp. 179–182, China, August 2008.

[74] G. I. Parisi, C. Weber, and S. Wermter, "Human action recognition with hierarchical growing neural gas learning," in *Proceeding of Artificial Neural Networks and Machine Learning on Springer – ICANN 2014*, vol. 8681 of *Lecture Notes in Computer Science*, pp. 89–96, Springer International Publishing, Switzerland, 2014.

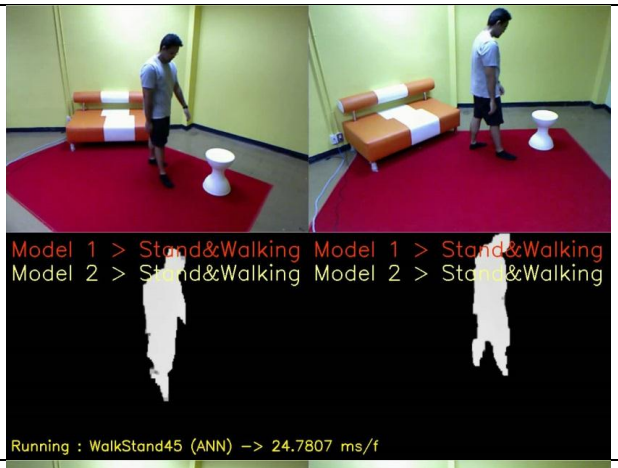
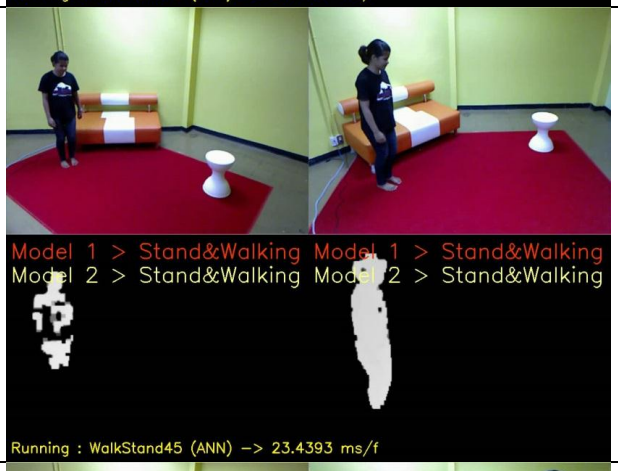
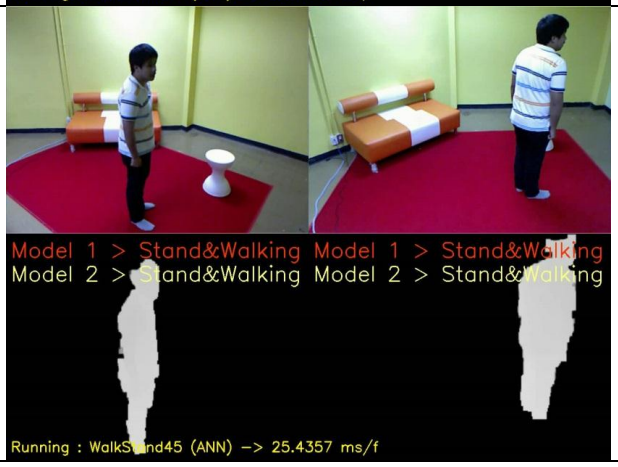
[75] N. Sawant and K. K. Biswas, "Human action recognition based on spatio-temporal features," in *Proceedings of the Pattern Recognition and Machine Intelligence Springer*, Lecture Notes in Computer Science, pp. 357–362, Springer, Heidelberg, Berlin, 2009.

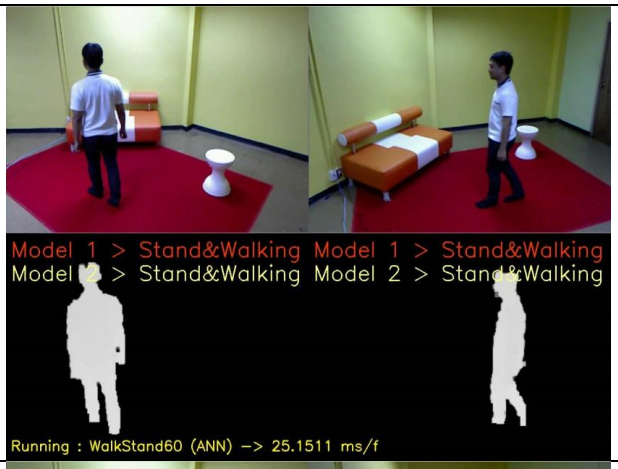
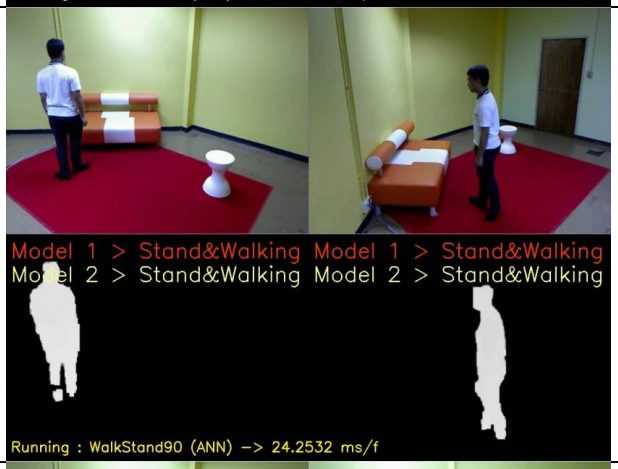
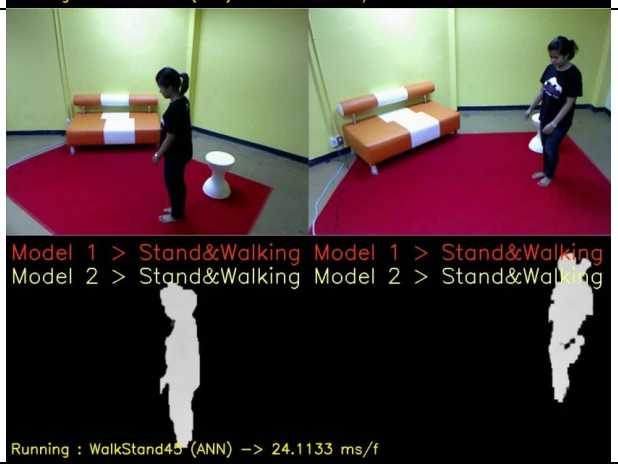
[76] N. Suvonvorn, *Prince of Songkla University (PSU) Multi-view profile-based action RGBD dataset*, 2017, http://fivedots.coe.psu.ac.th/~kom/?p=1483 (accessed 20 December 2017).

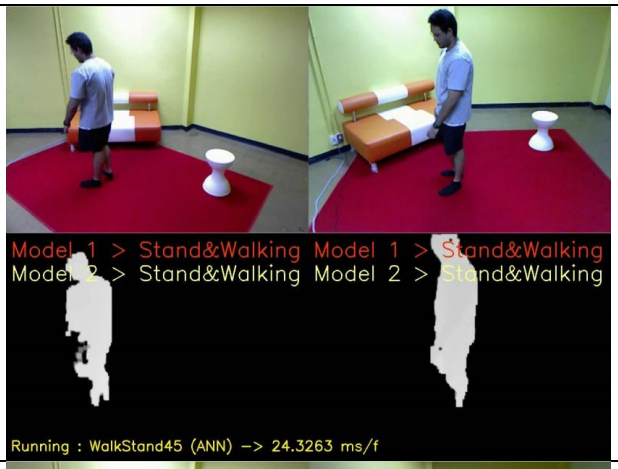
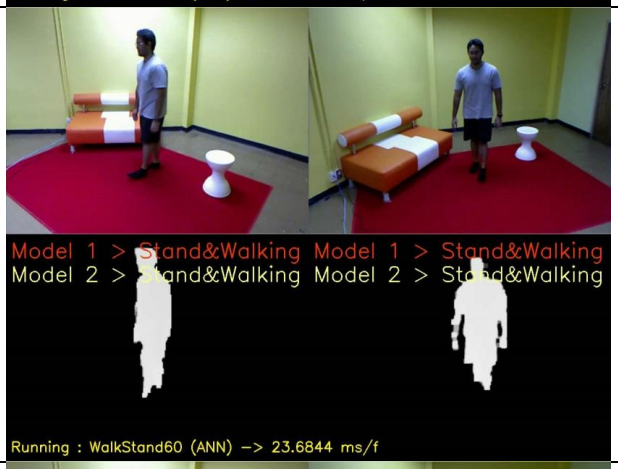
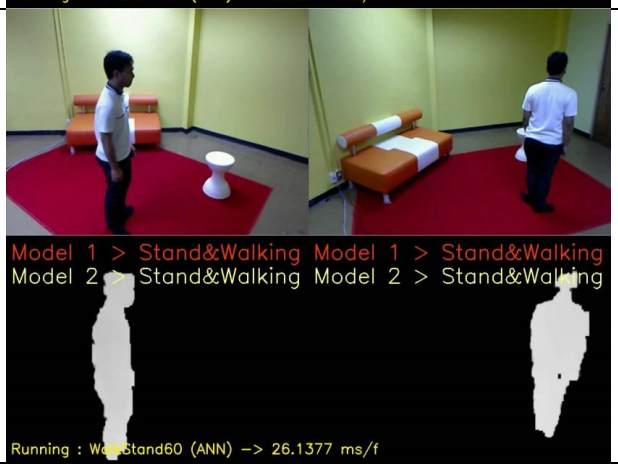
[77] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3DPost multi-view and 3D human action/interaction database," in *Proceeding of the 6th European Conference for Visual Media Production (CVMP '09)*, pp. 159–168, London, UK, November 2009.

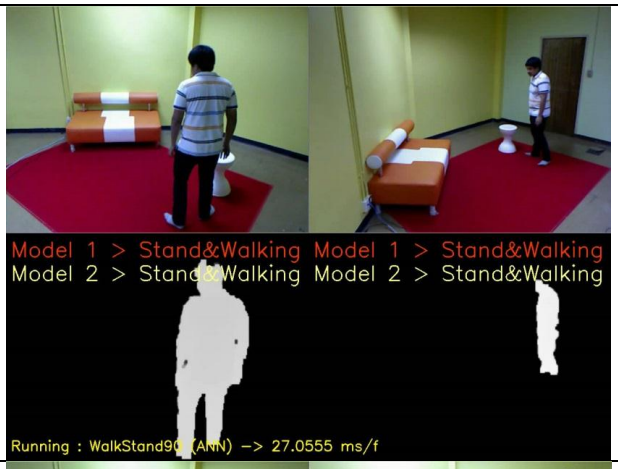
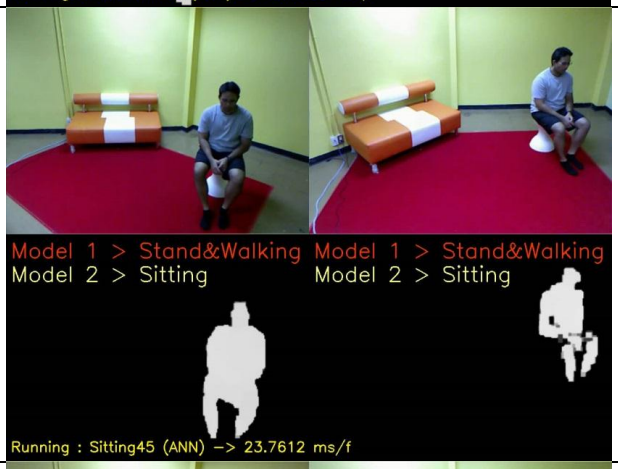
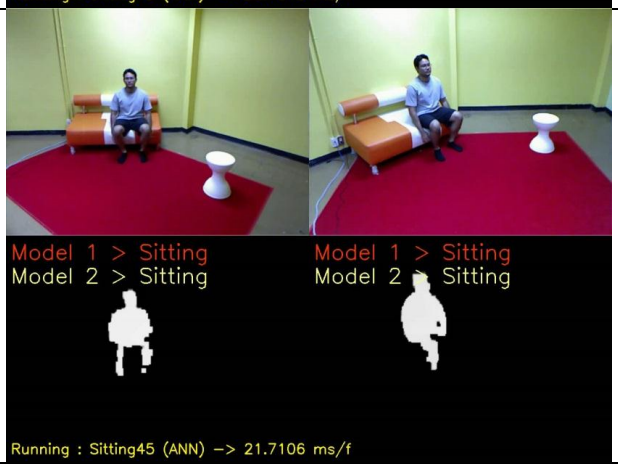
ภาคผนวก ง.
ตัวอย่างผลการทดสอบการรู้จำท่าทางระดับการฟิวชันฟีเจอร์ในระดับล่างจากหลายมุมมอง (เพิ่มเติม)

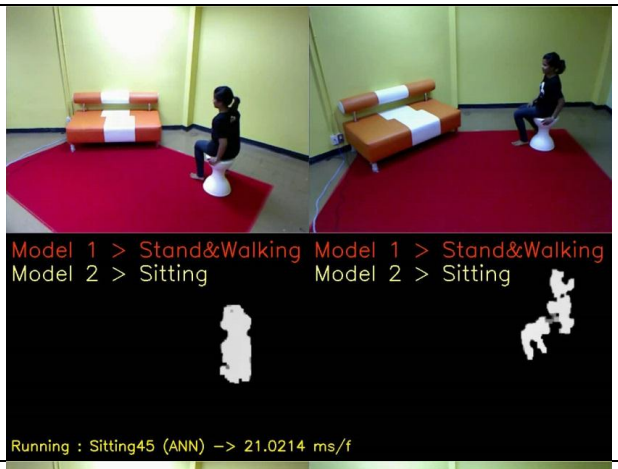
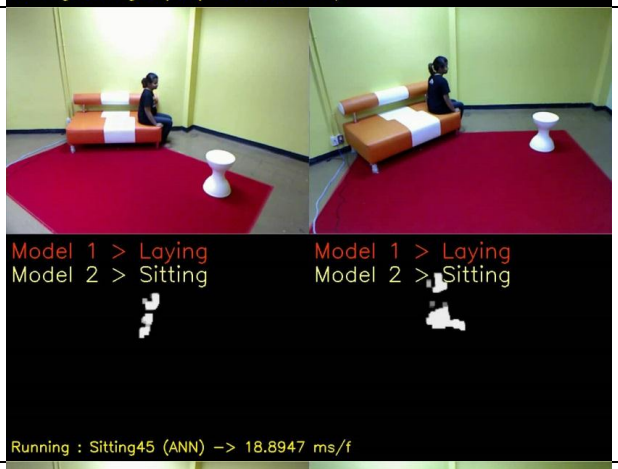
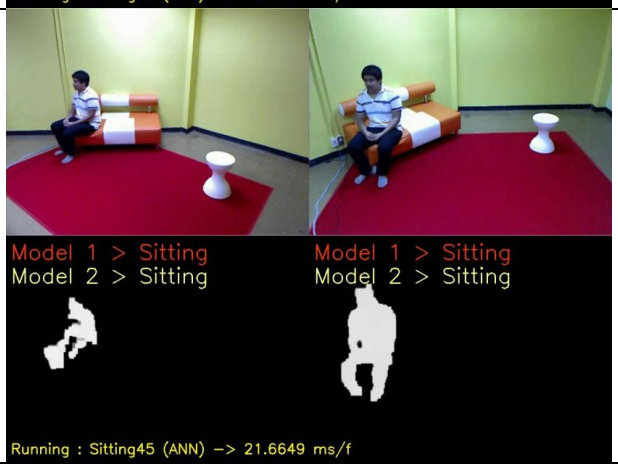
ตารางที่ ง-1 ตัวอย่างการทดสอบการรู้จำท่าทางระดับการฟิวชันฟีเจอร์ในระดับล่างในชุดข้อมูล PSU

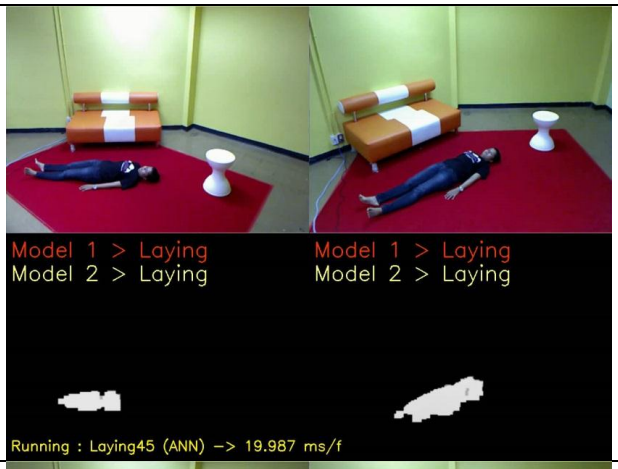
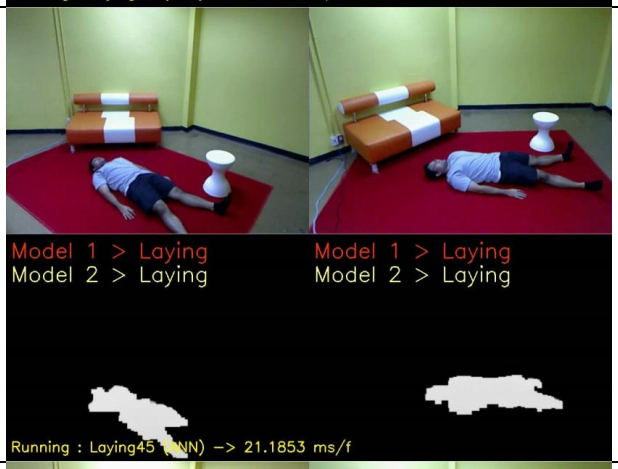
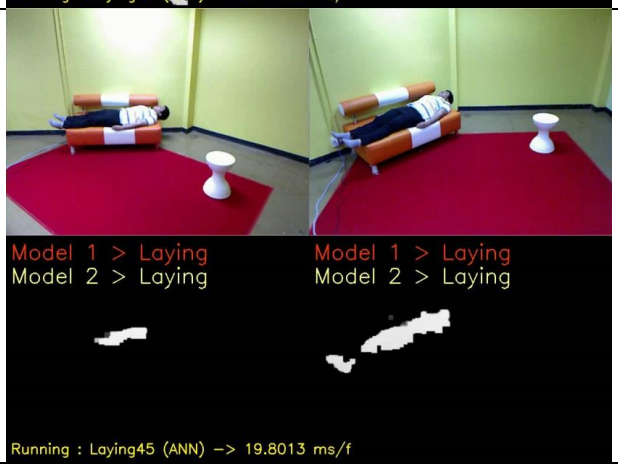
ท่าทาง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน	
ยืนและเดิน	
ยืนและเดิน	

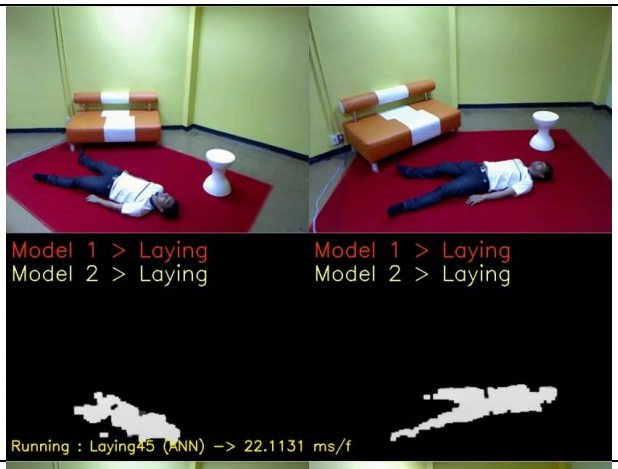
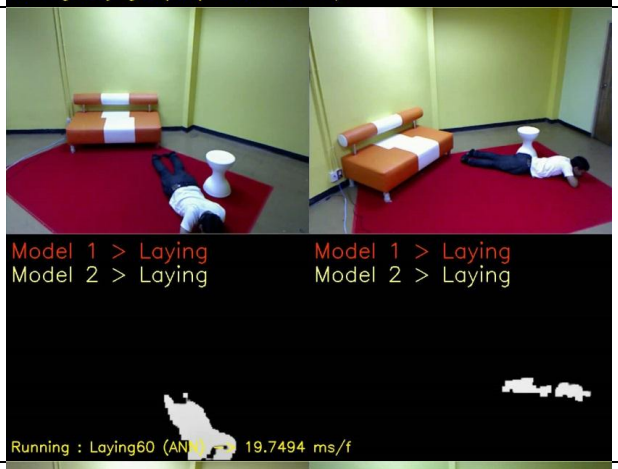
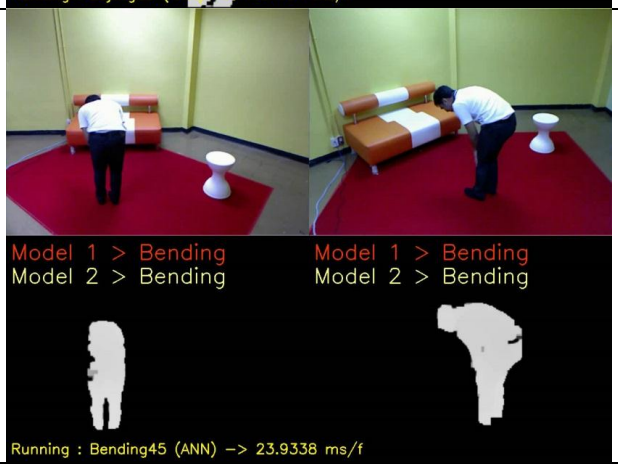
ท่าทาง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน	
ยืนและเดิน	
ยืนและเดิน	

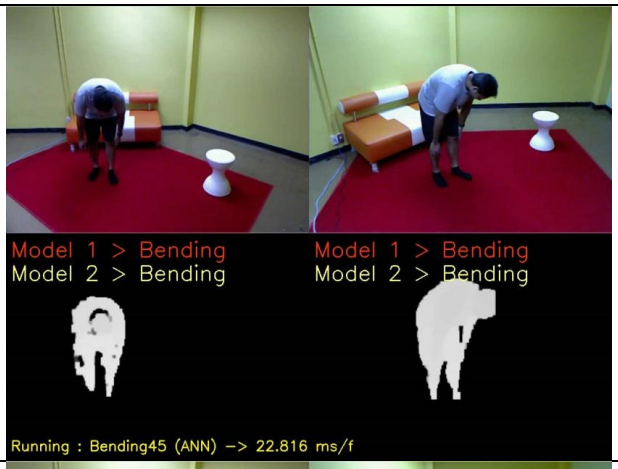
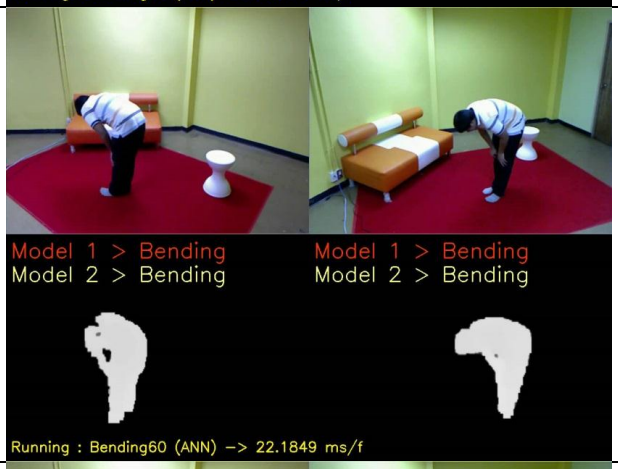
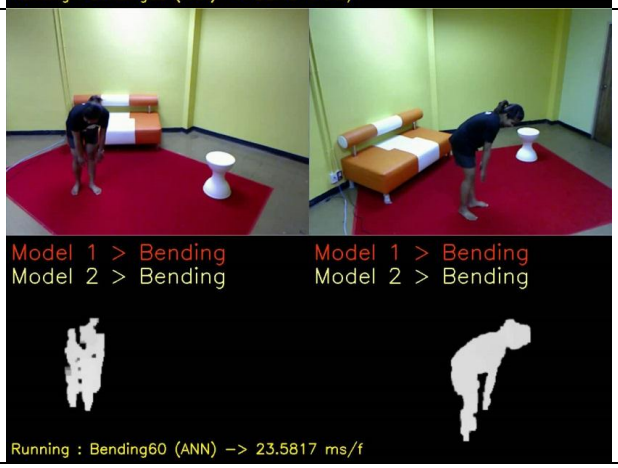
ท่าทาง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน	
ยืนและเดิน	
ยืนและเดิน	

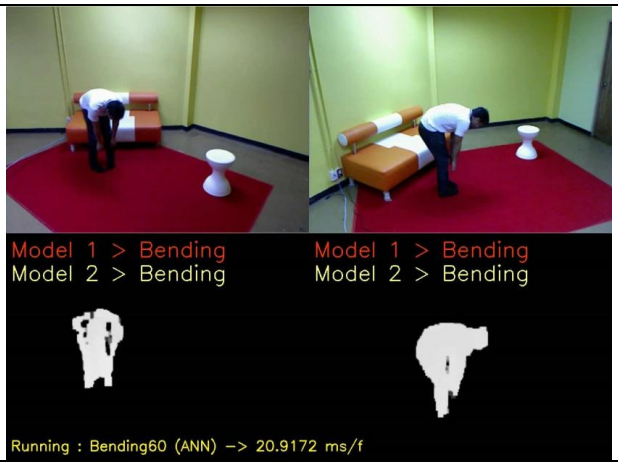
ท่าทาง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน	
นั่ง	
นั่ง	

ท่าทาง	ภาพตัวอย่างการทดสอบ
นั่ง	
นั่ง	
นั่ง	

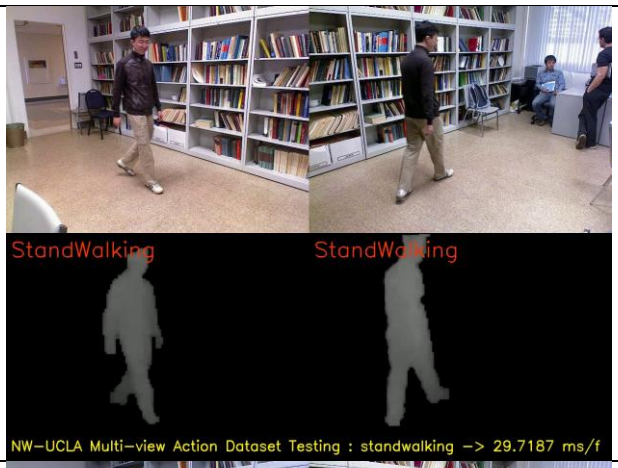
ท่าทาง	ภาพตัวอย่างการทดสอบ
นอน	
นอน	
นอน	

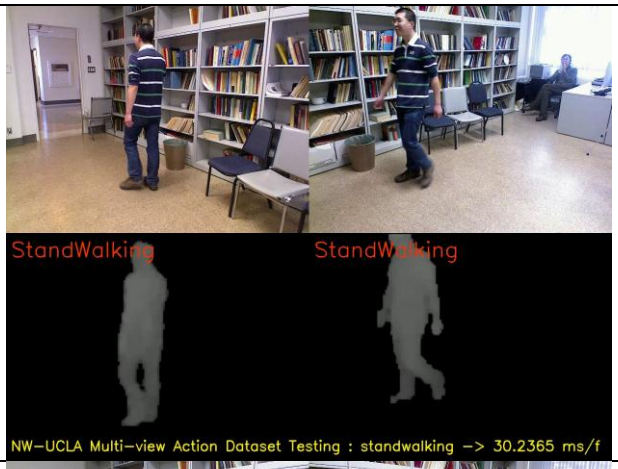
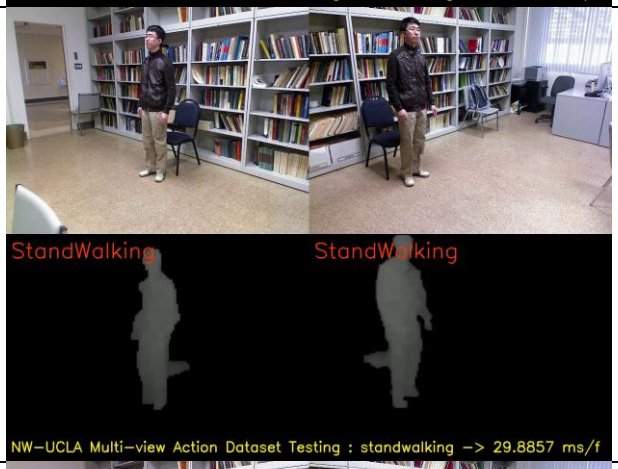
ท่าทาง	ภาพตัวอย่างการทดสอบ
นอน	
นอน	
ก้ม	

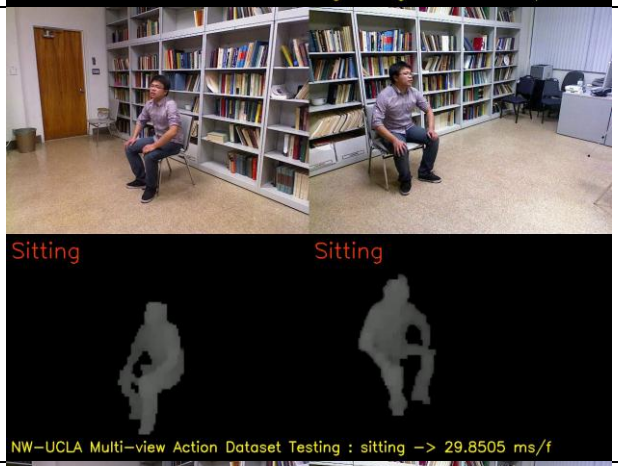
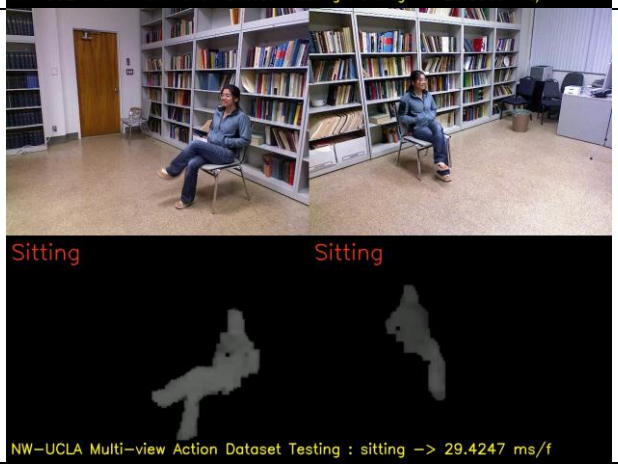
ท่าทาง	ภาพตัวอย่างการทดสอบ
ก้ม	
ก้ม	
ก้ม	

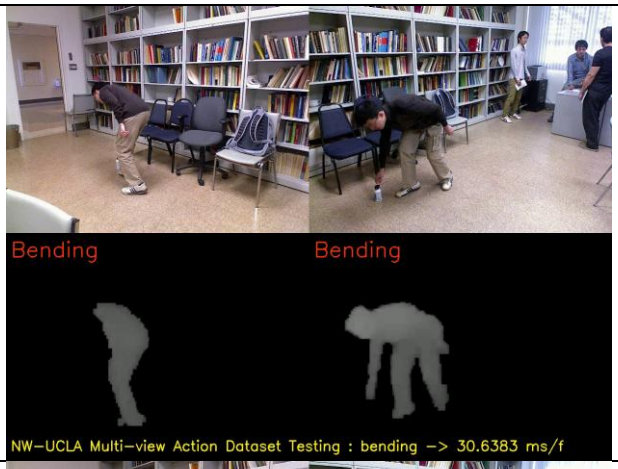
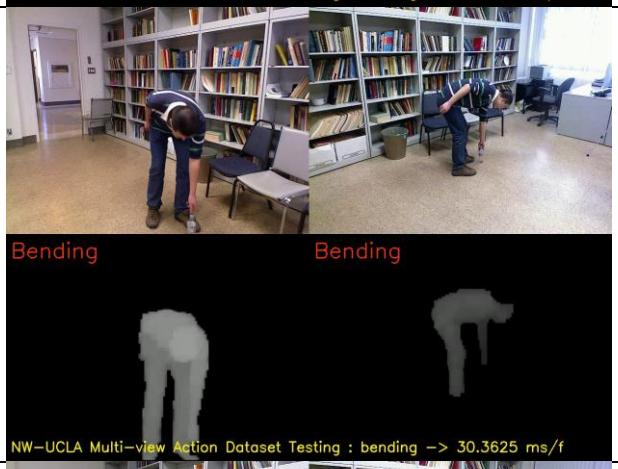
ท่าทาง	ภาพตัวอย่างการทดสอบ
ก้ม	

ตารางที่ ง-2 ตัวอย่างการทดสอบการรู้จำท่าทางระดับการฟิวชันฟีเจอร์ในระดับล่างในชุดข้อมูล NW-UCLA

ท่าทาง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน	
ยืนและเดิน	

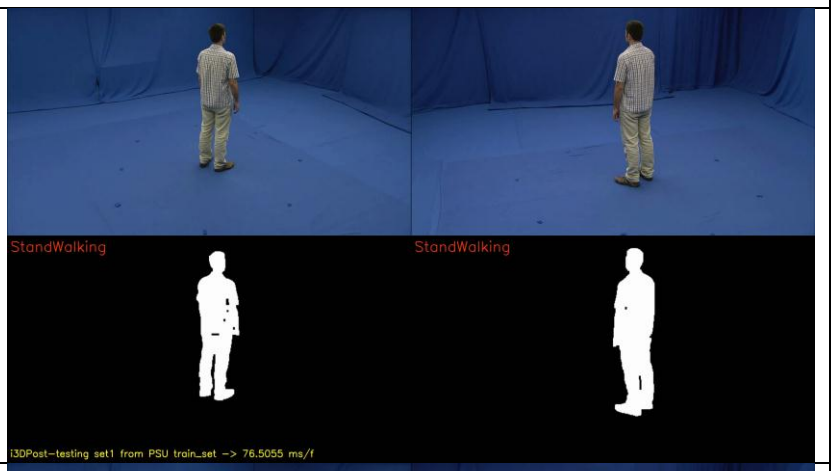
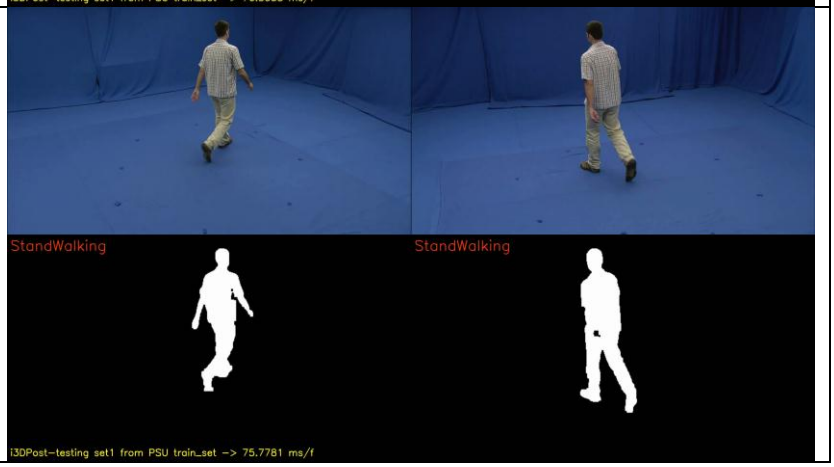
ท่าทาง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน	
ยืนและเดิน	
นั่ง	

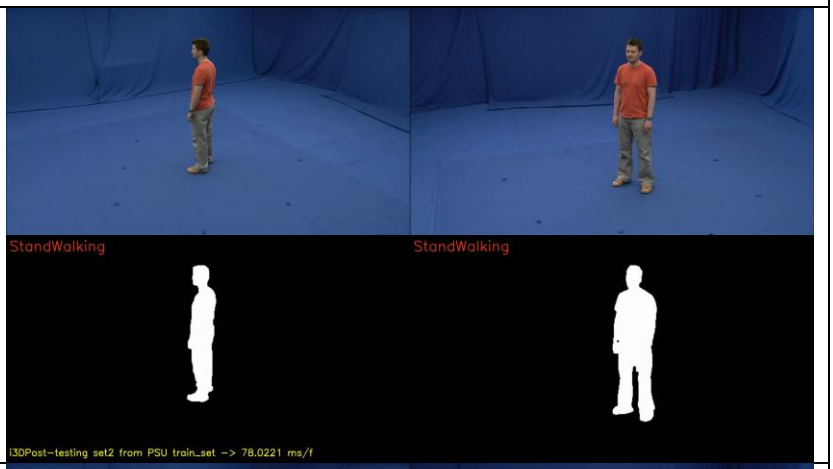
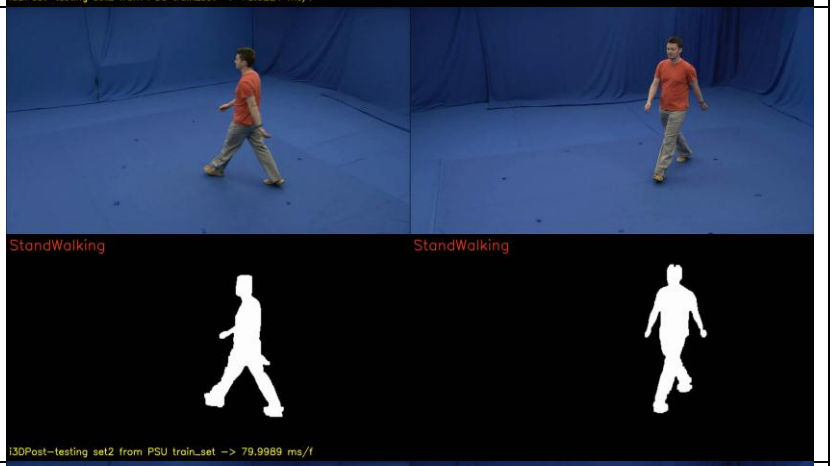
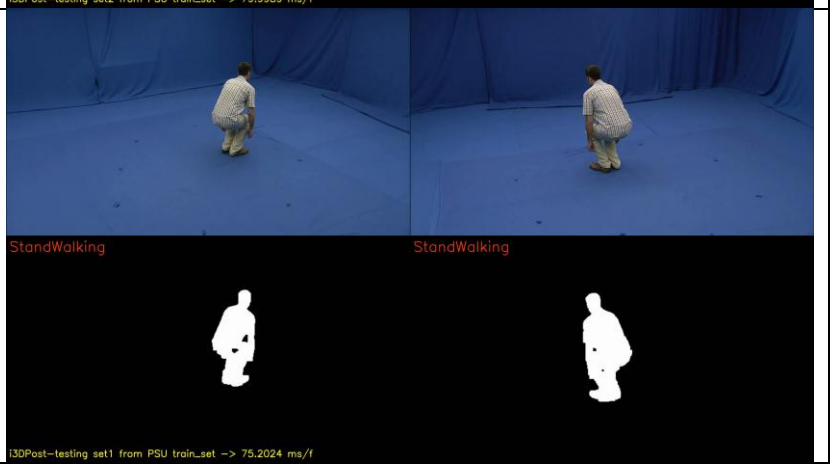
ท่าทาง	ภาพตัวอย่างการทดสอบ
นั่ง	
นั่ง	
นั่ง	

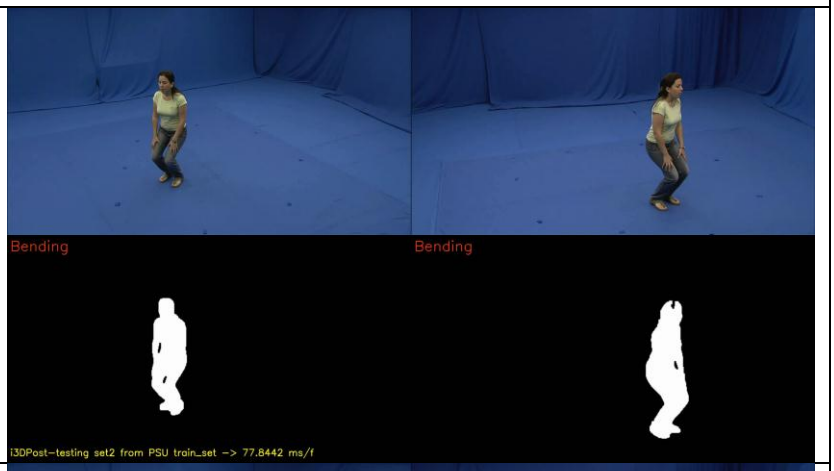
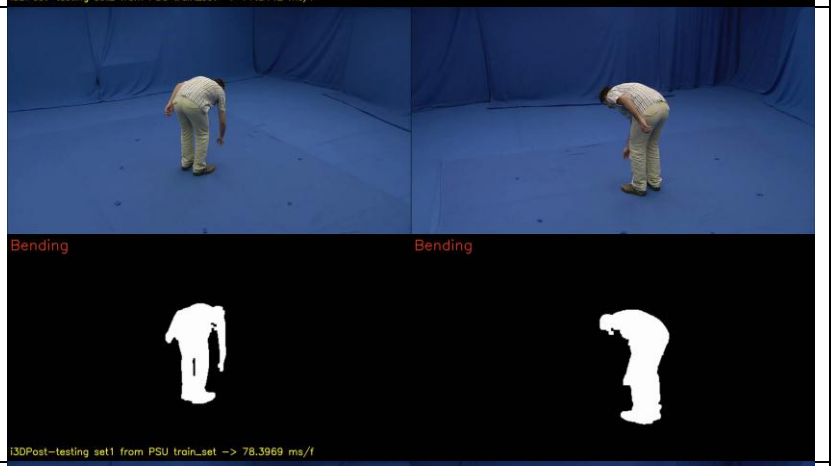
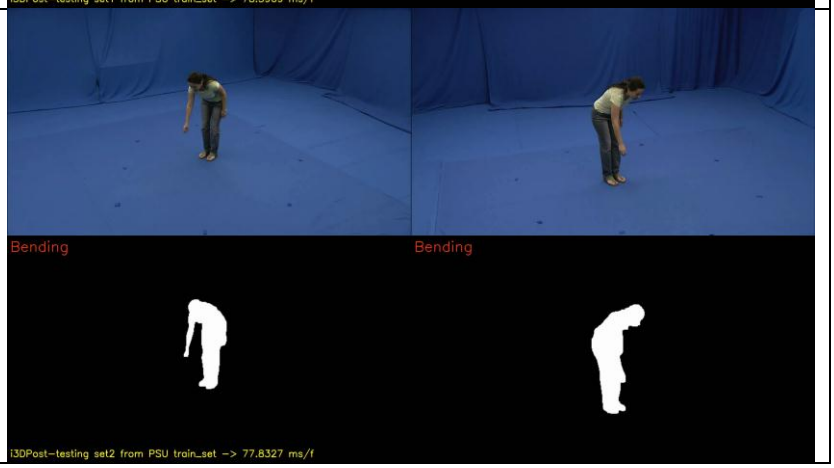
ท่าทาง	ภาพตัวอย่างการทดสอบ
ก้ม	
ก้ม	
ก้ม	

ท่าทาง	ภาพตัวอย่างการทดสอบ
ก้ม	

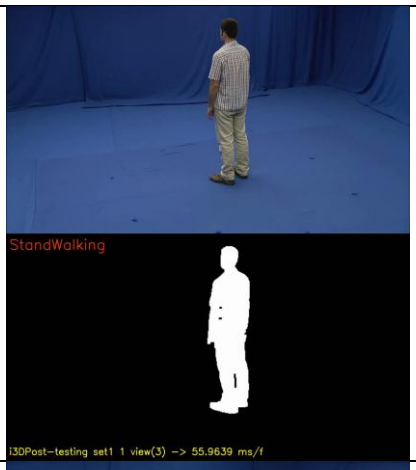
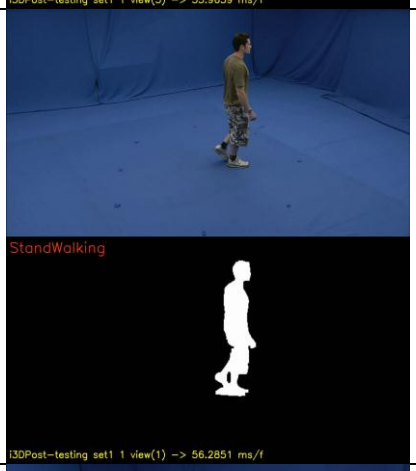
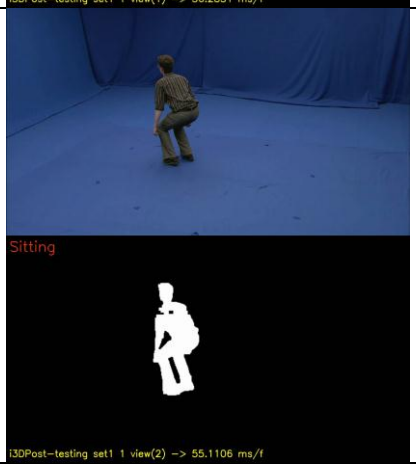
ตารางที่ ง-3 ตัวอย่างการทดสอบการรู้จำท่าทางระดับการฟิวชันฟีเจอร์ในระดับล่างในชุดข้อมูล i3DPost โดยใช้แบบจำลองที่สอนจากชุดข้อมูล PSU

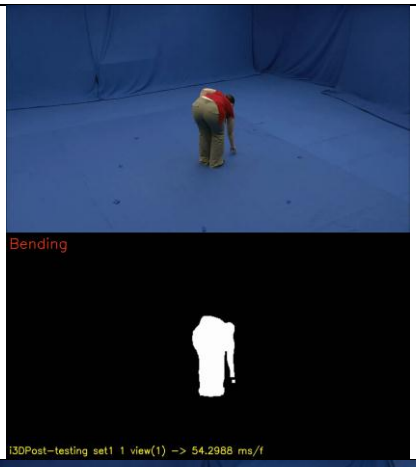
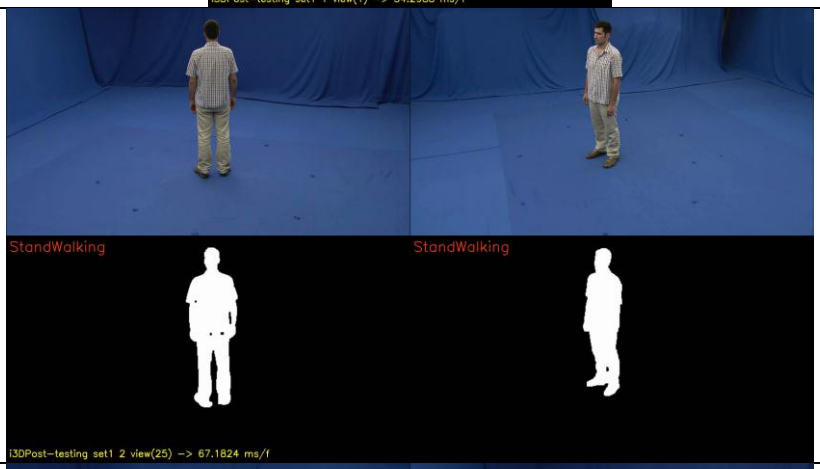
ท่าทาง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน	
ยืนและเดิน	

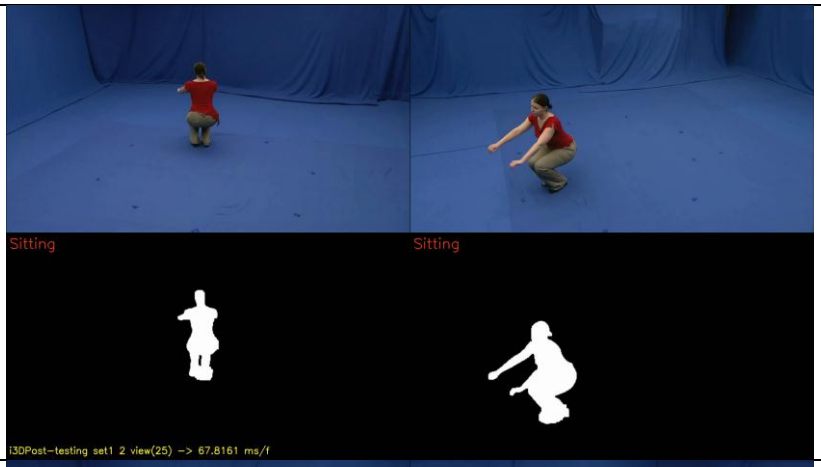
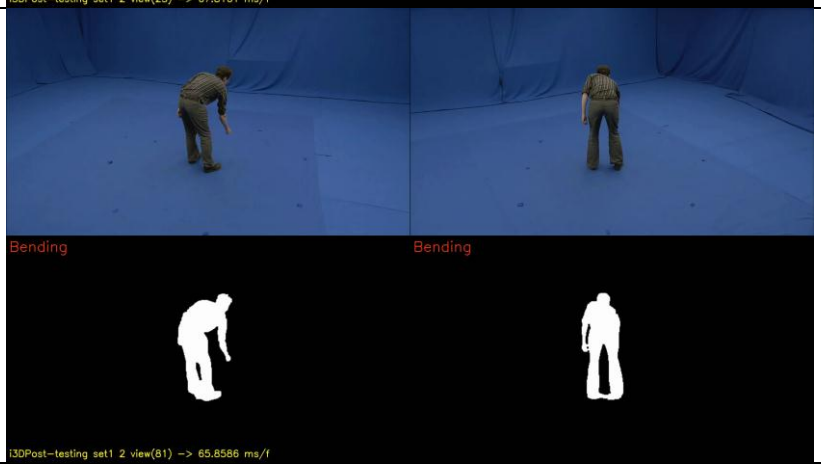
ท่าทาง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน	
ยืนและเดิน	
นั่ง	

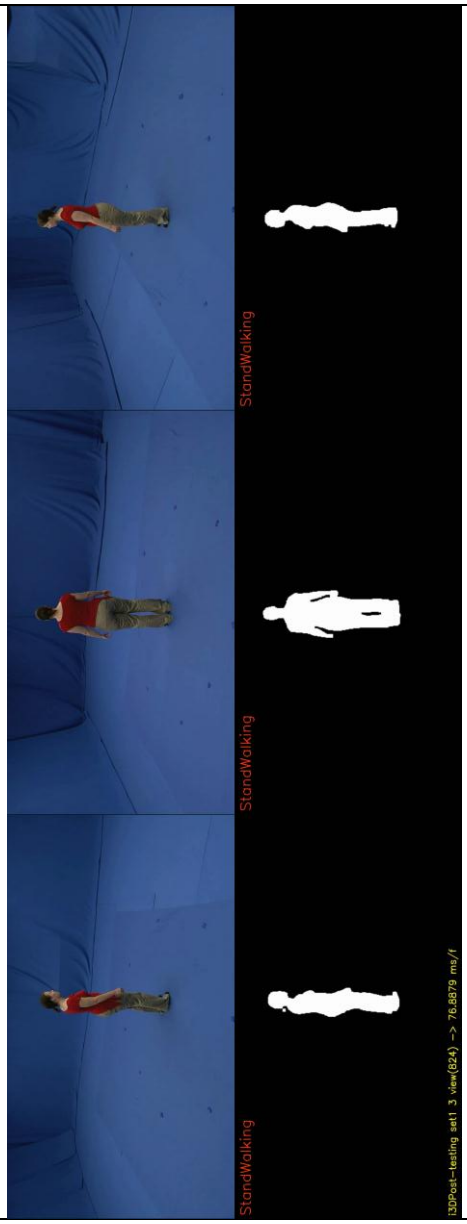
ท่าทาง	ภาพตัวอย่างการทดสอบ
นั่ง	
ก้ม	
ก้ม	

ตารางที่ ง-4 ตัวอย่างการทดสอบการรู้จำท่าทางระดับการฟิวชันฟีเจอร์ในระดับล่างในชุดข้อมูล i3DPost โดยใช้แบบจำลองที่สอนชุดใหม่

ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน / 1	
ยืนและเดิน / 1	
นั่ง / 1	

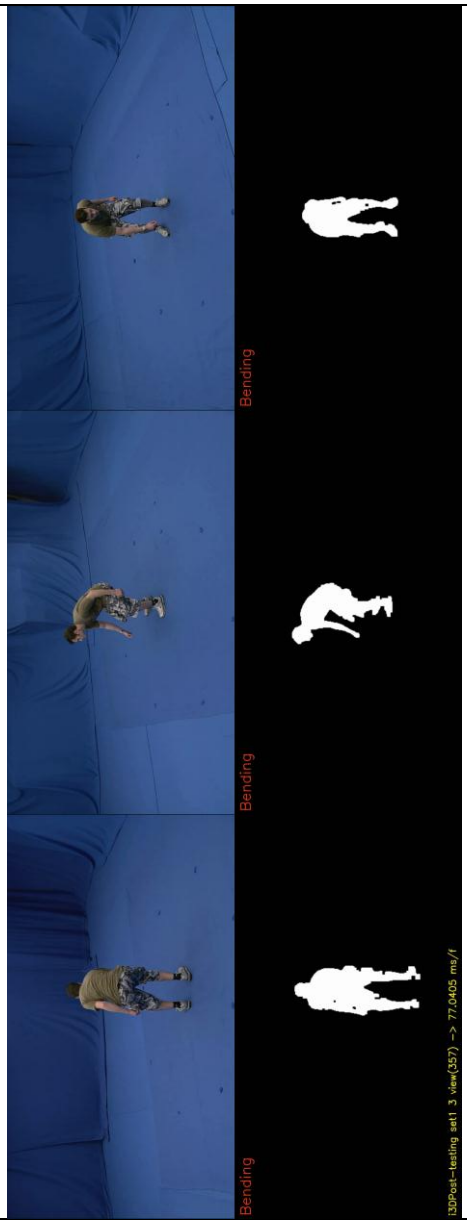
ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
ก้ม / 1	
ยืนและเดิน / 2	
ยืนและเดิน / 2	

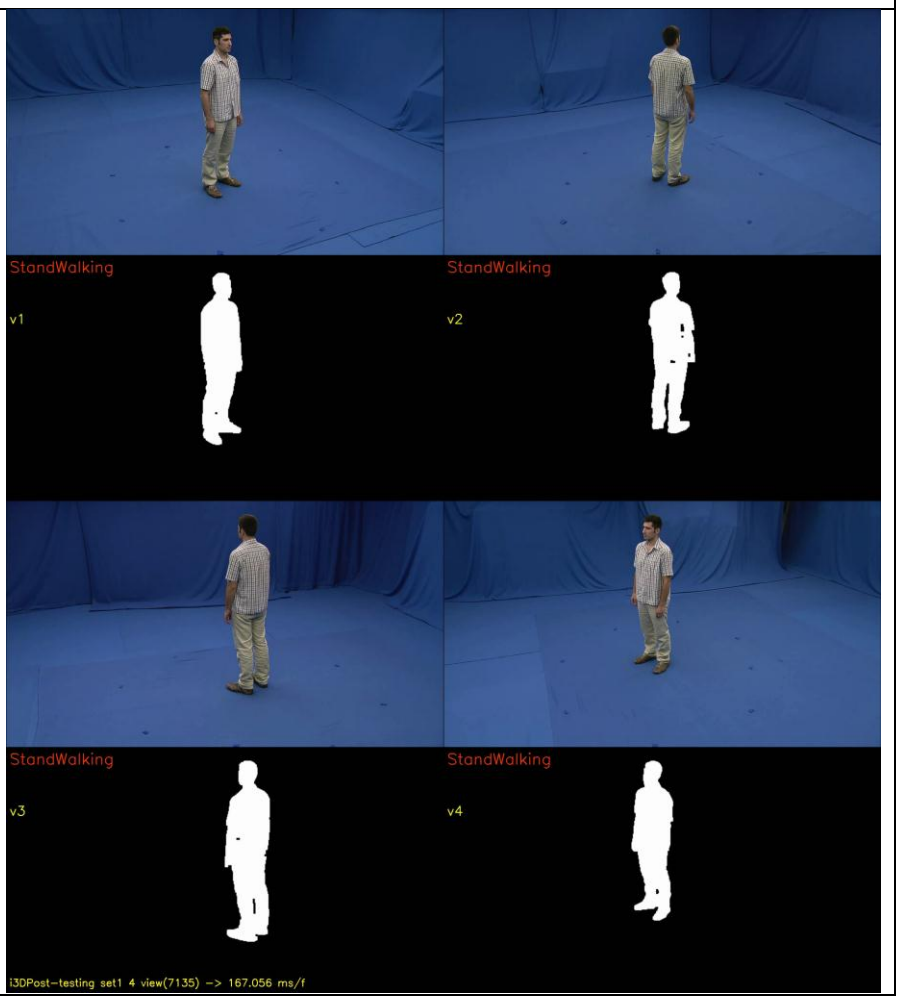
ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
นั่ง / 2	
ก้ม / 2	

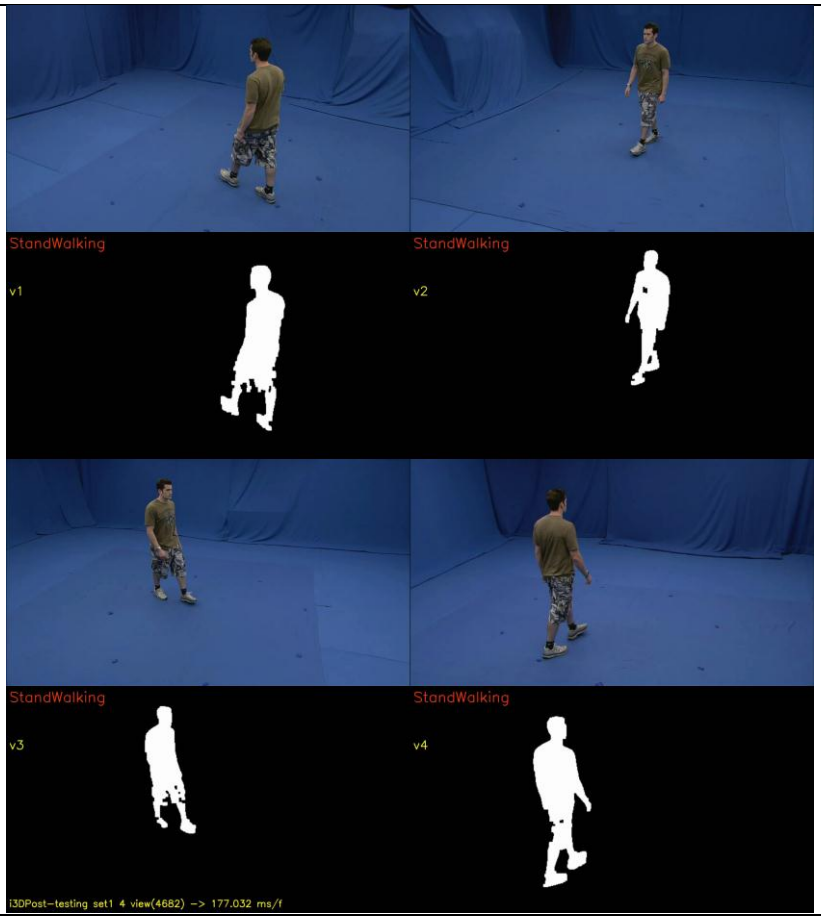
ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน / 3	

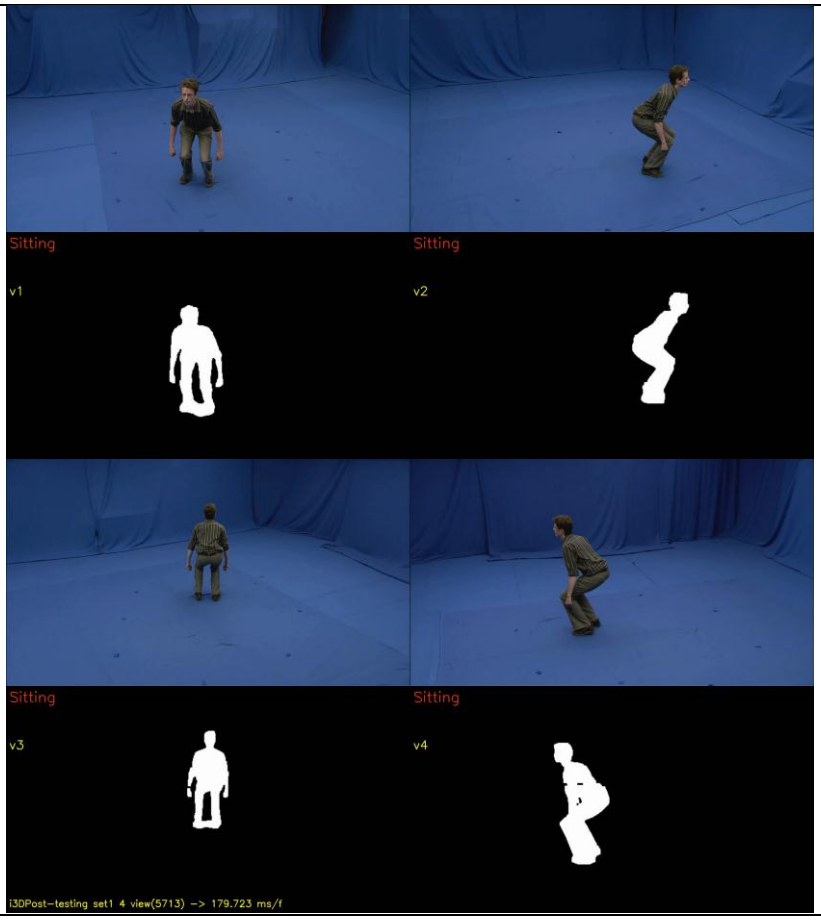
ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน / 3	

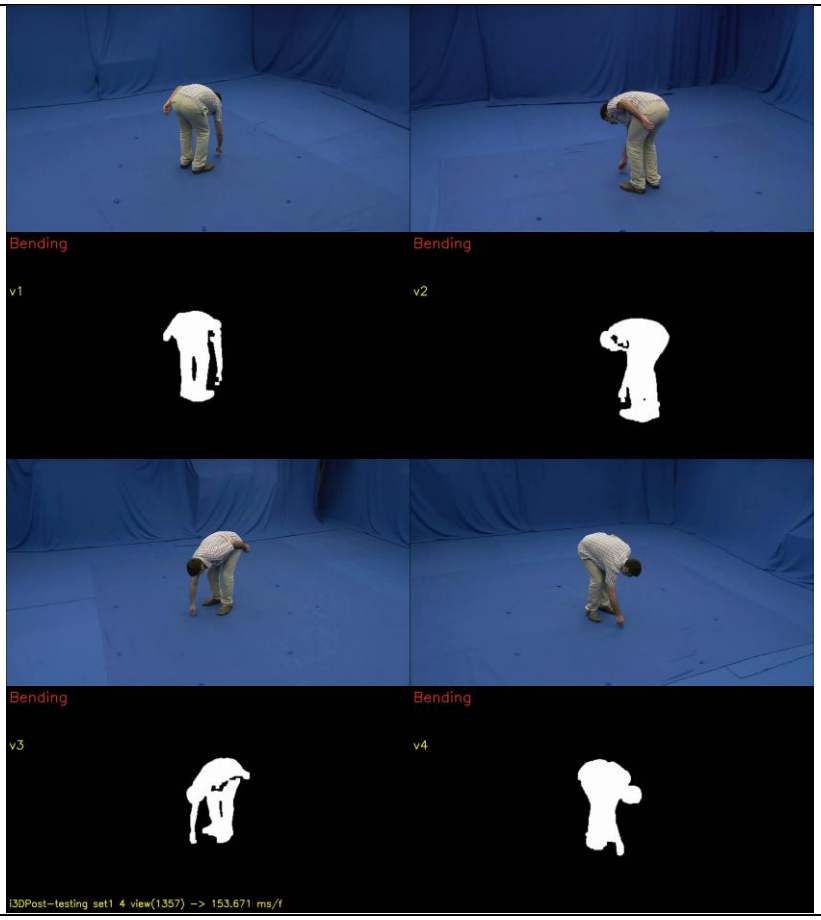
ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
นั่ง / 3	

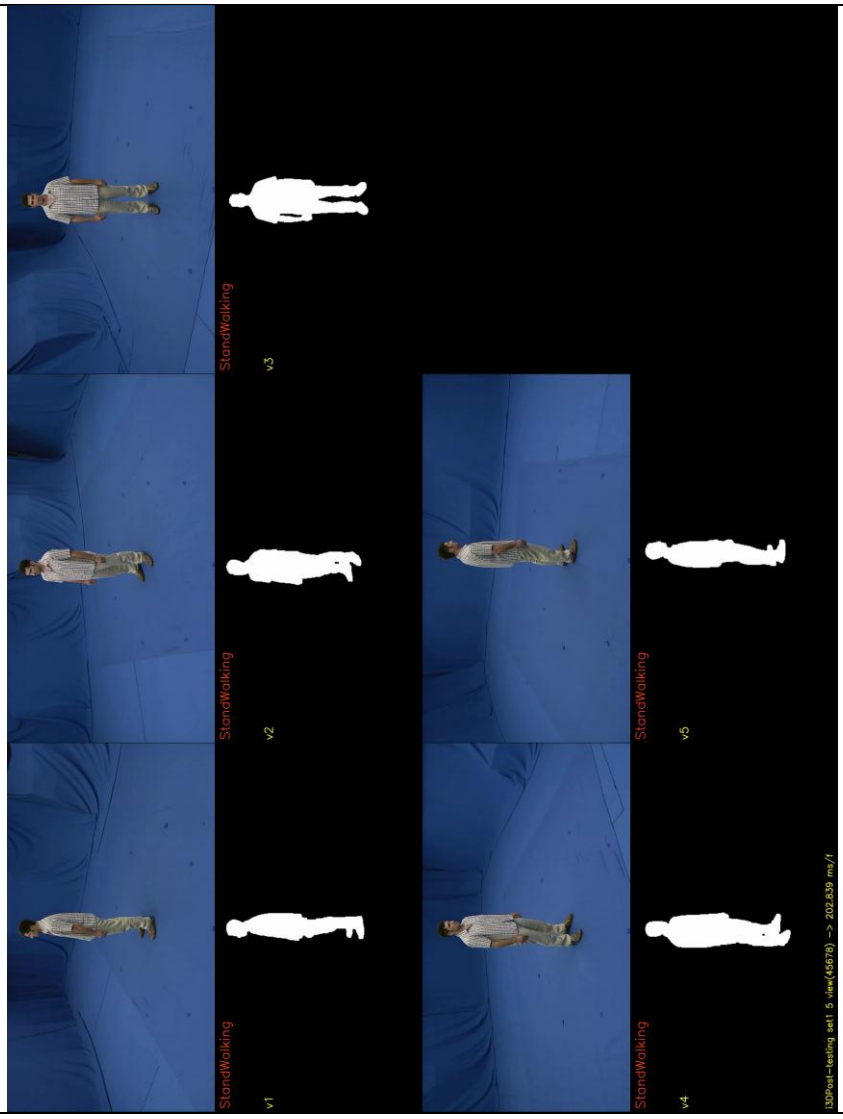
ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
ก้ม / 3	

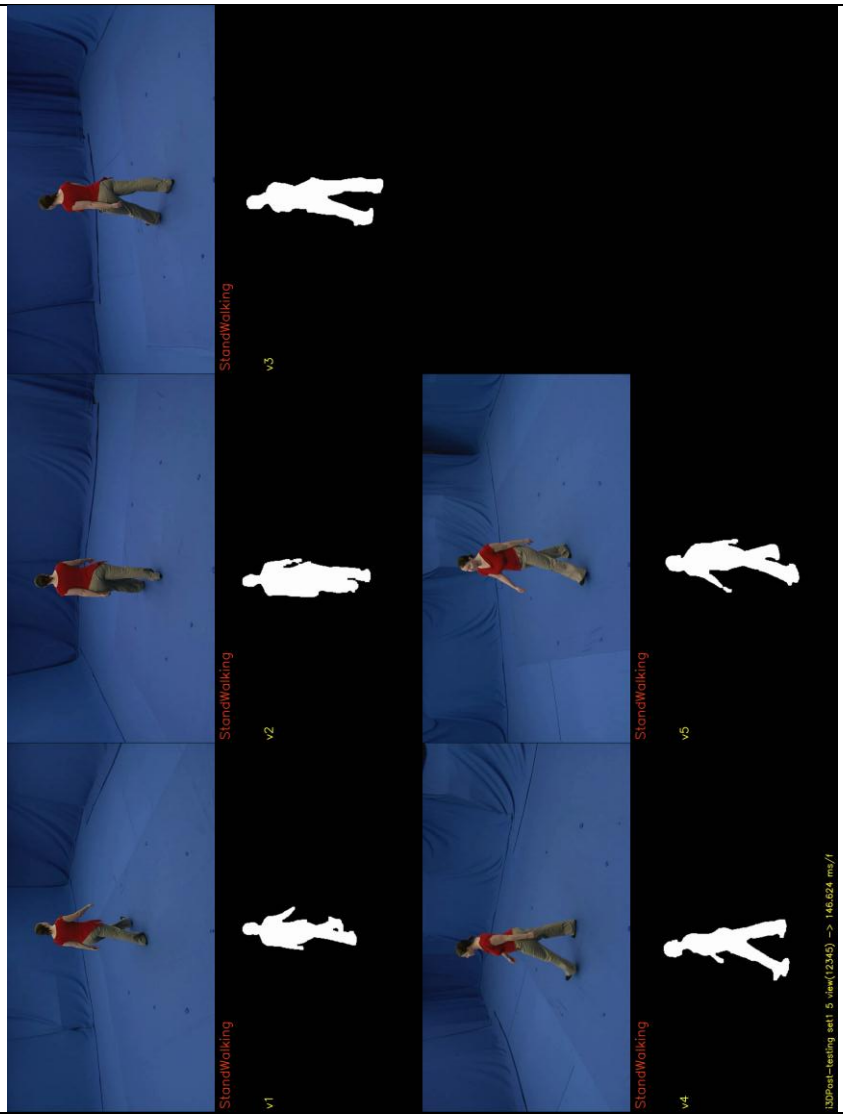
ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน / 4	

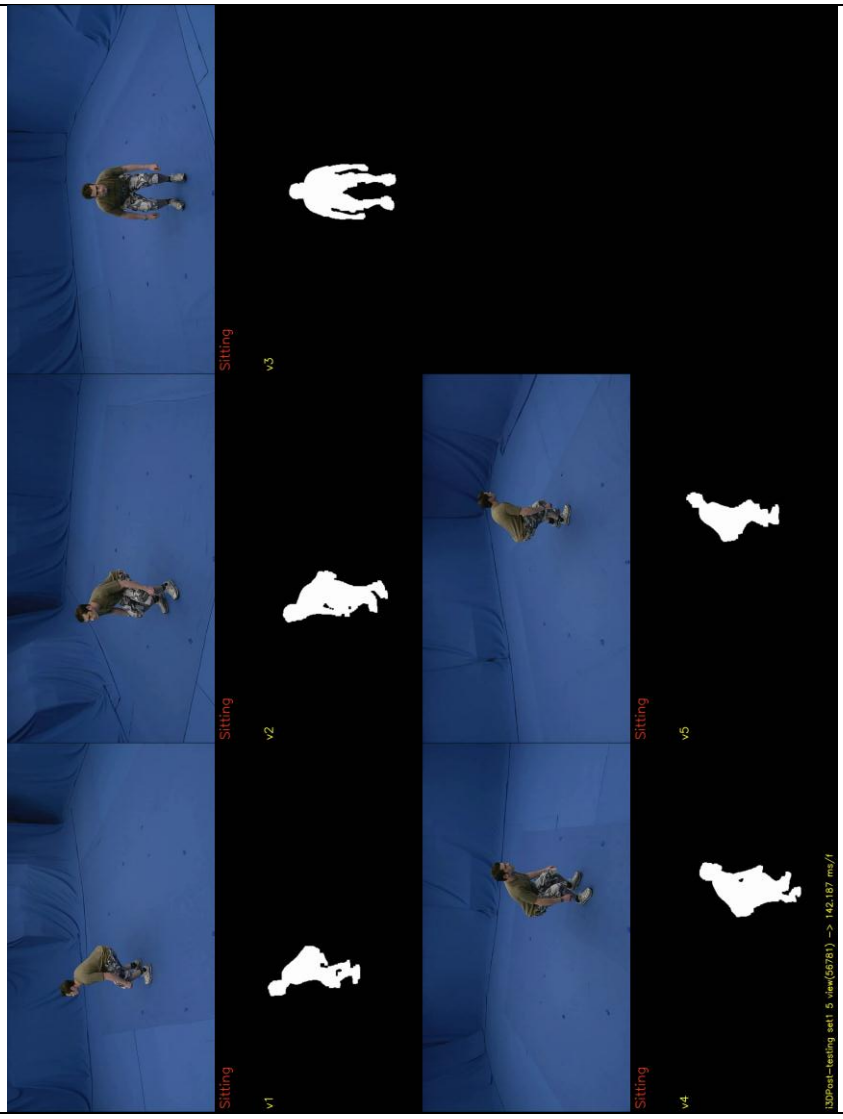
ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน / 4	

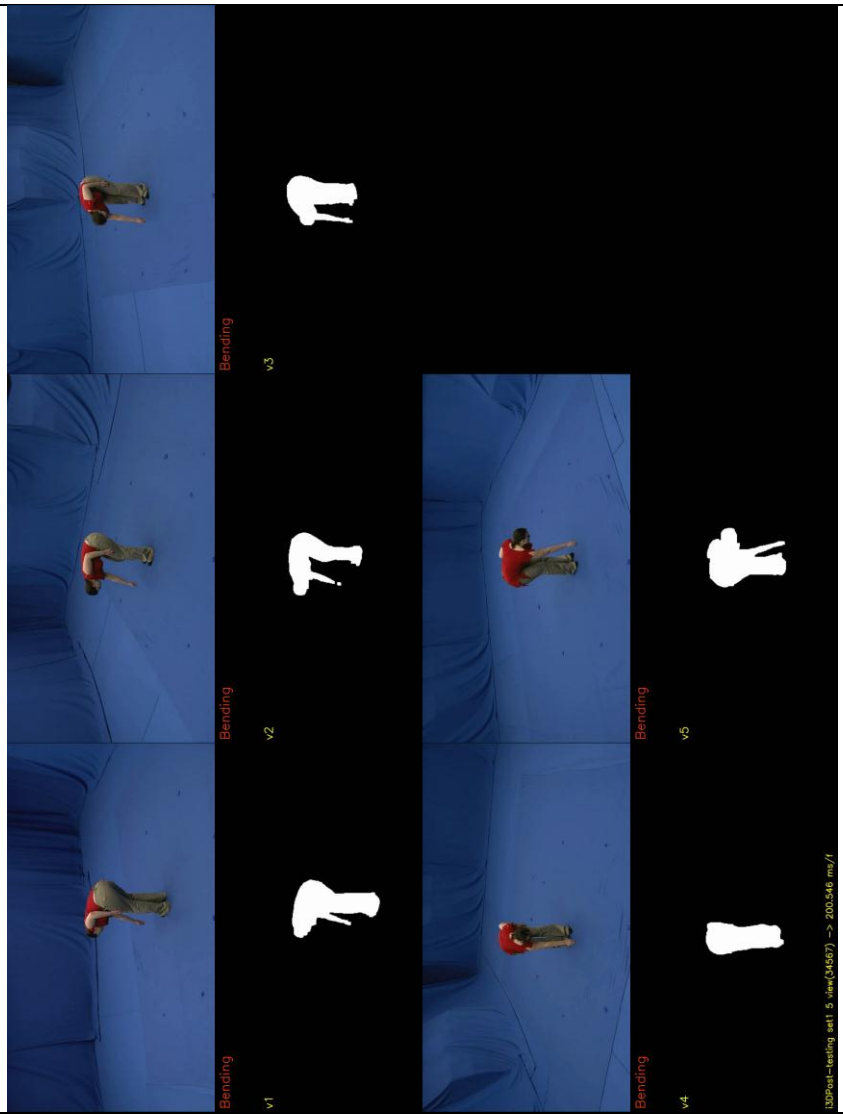
ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
นั่ง / 4	

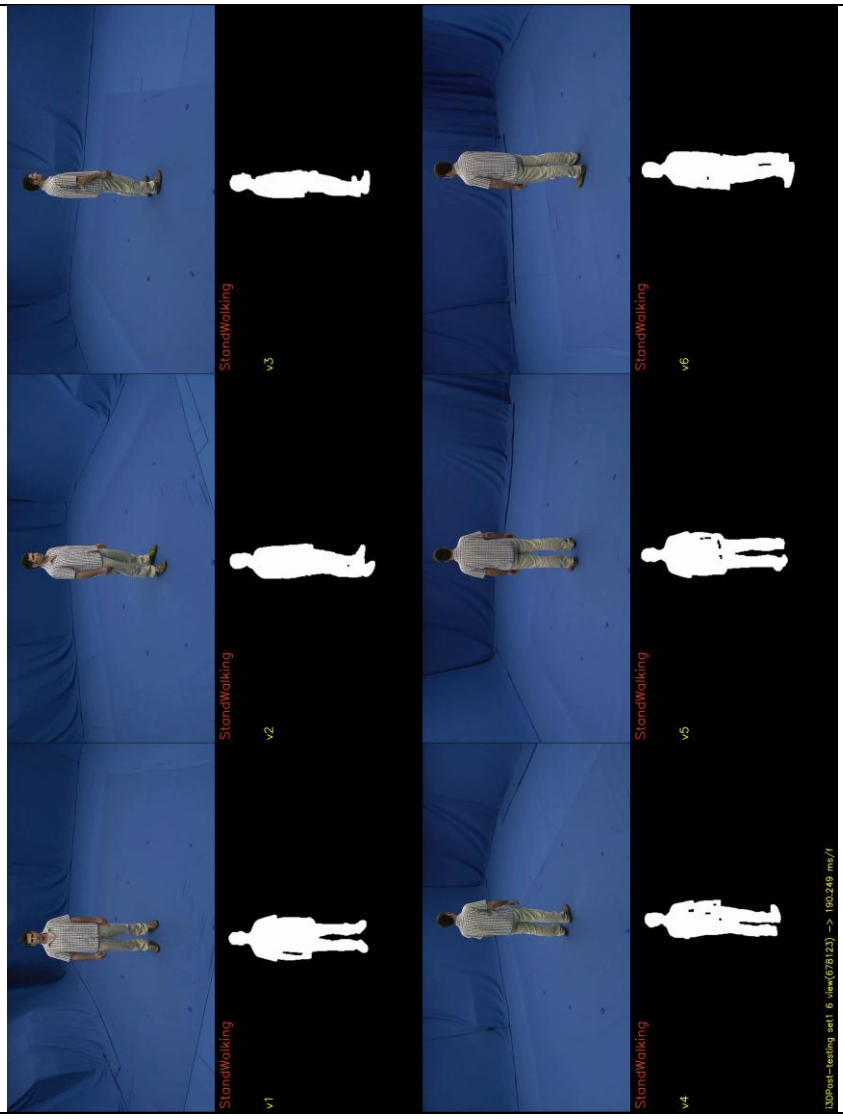
ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
ก้ม / 4	

ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน / 5	

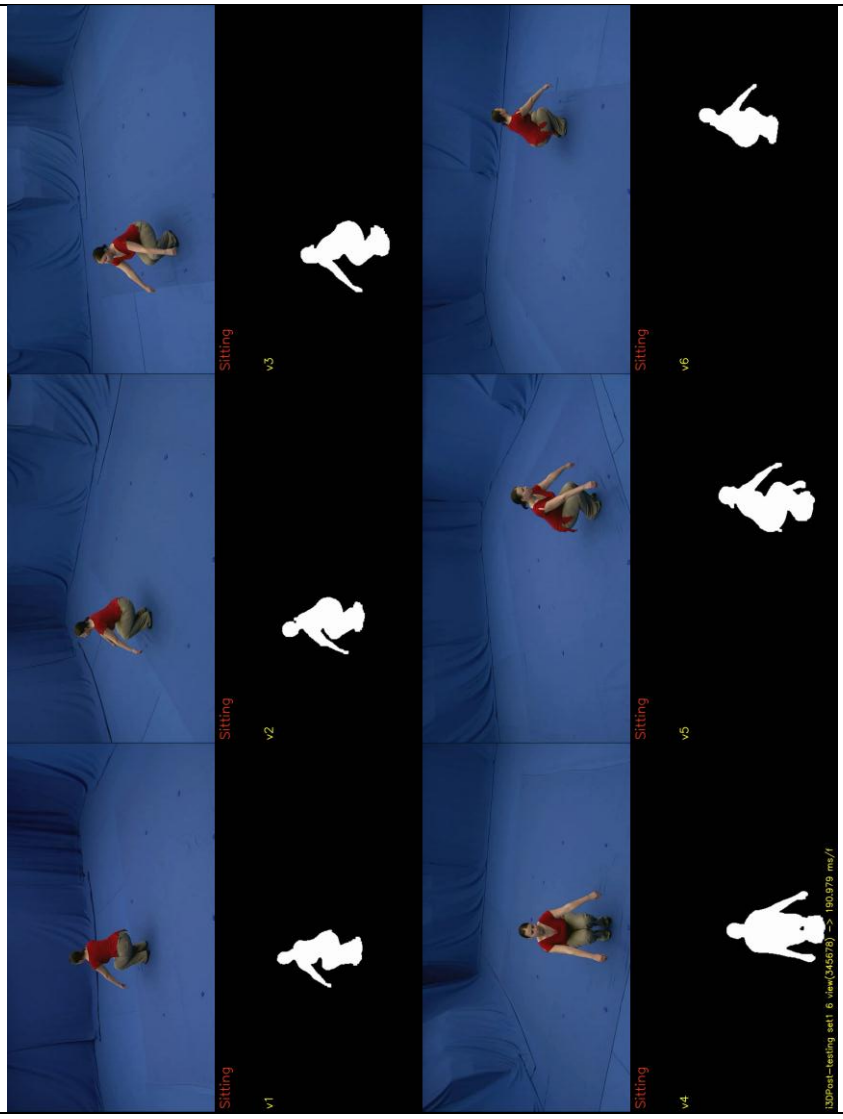
ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน / 5	

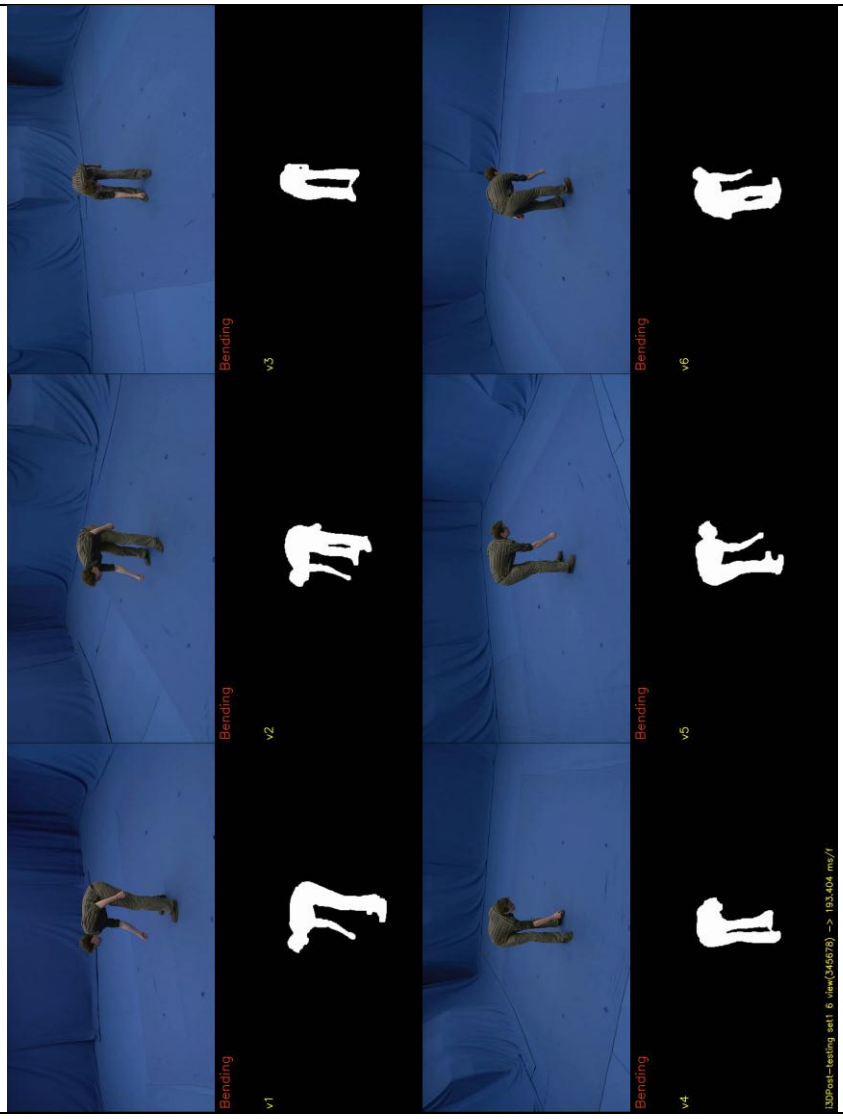
ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
นั่ง / 5	

ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
ก้ม / 5	

ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน / 6	

ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
ยืนและเดิน / 6	

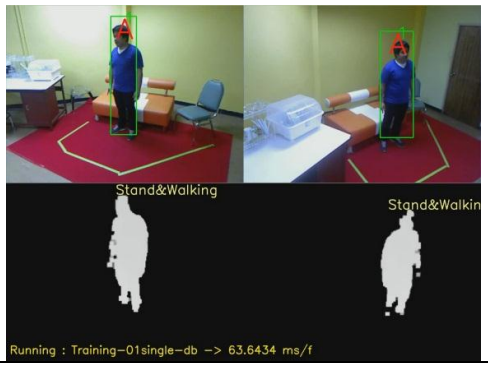
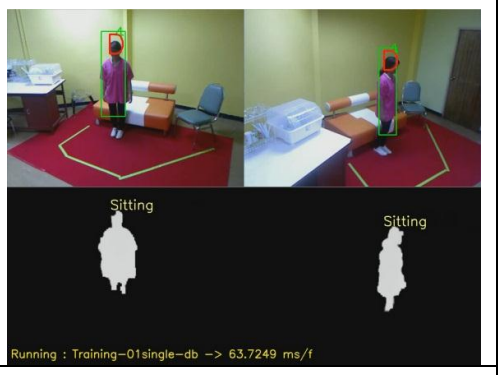
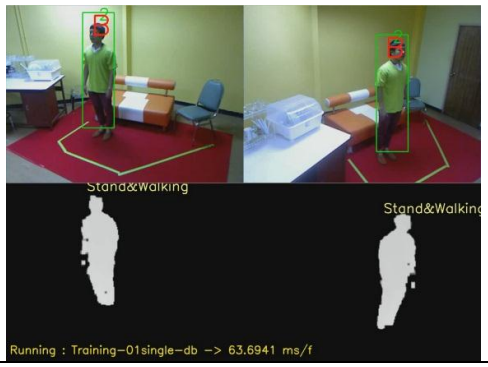
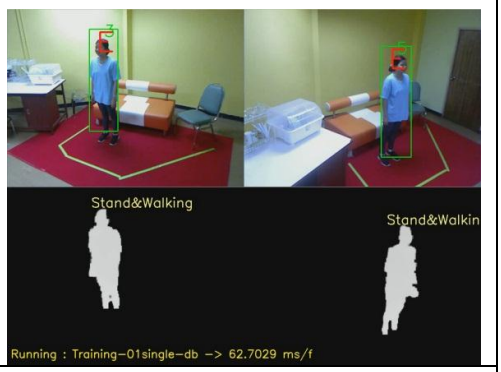
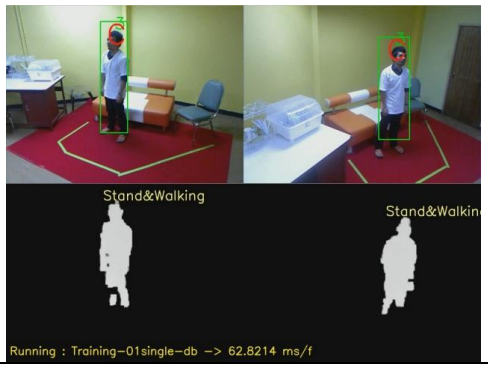
ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
นั่ง / 6	

ท่าทาง / จำนวนมุมมอง	ภาพตัวอย่างการทดสอบ
ก้ม / 6	

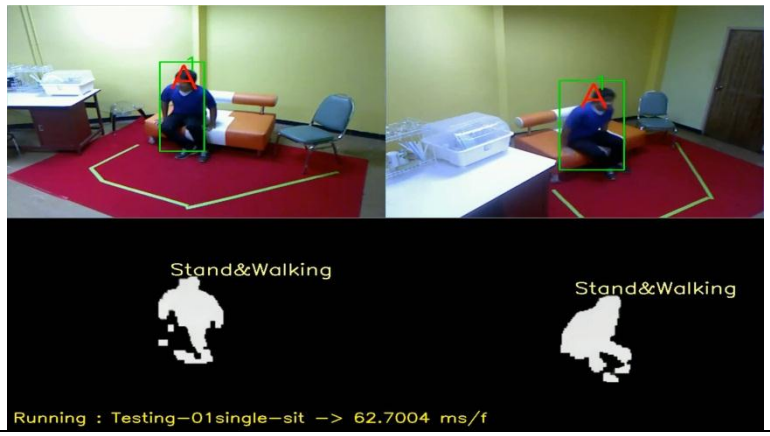
ภาคผนวก จ.

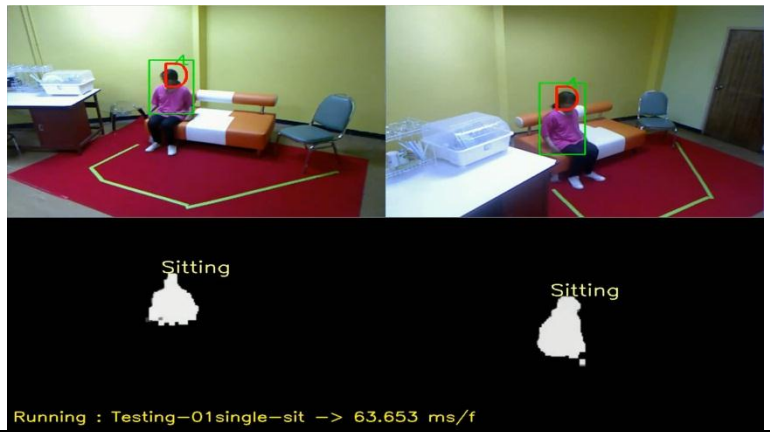
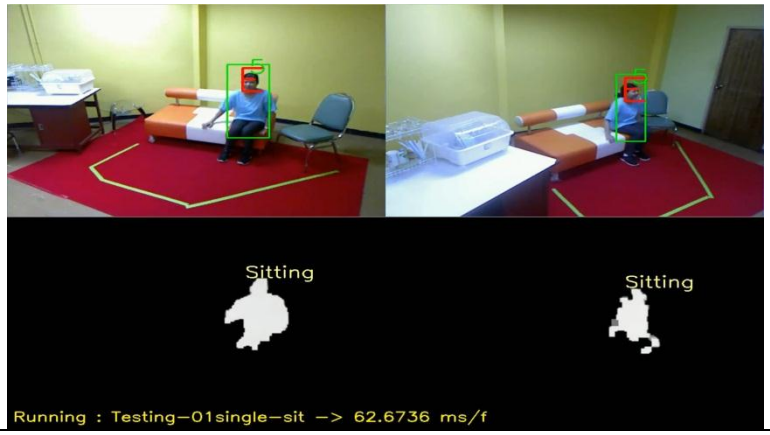
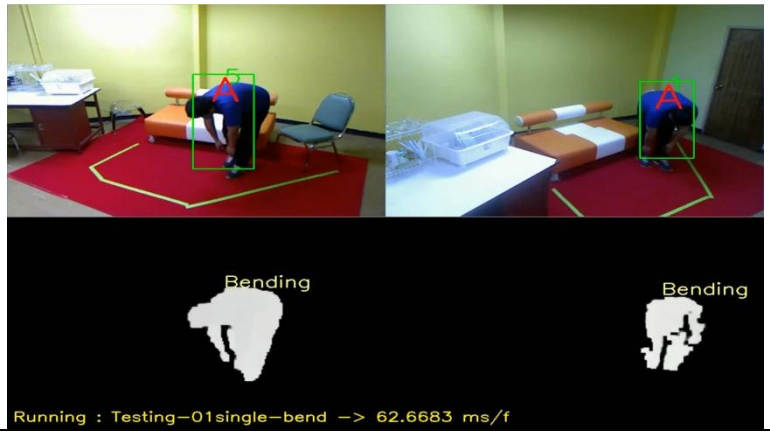
ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง (เพิ่มเติม)

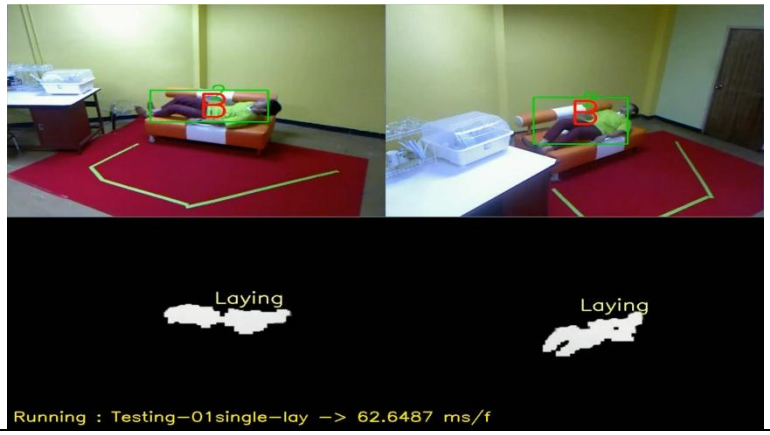
ตารางที่ จ-1 ตัวอย่างบุคคลที่เข้าทดสอบครั้งละหนึ่งคนและ Global ID สำหรับ Dataset # 1

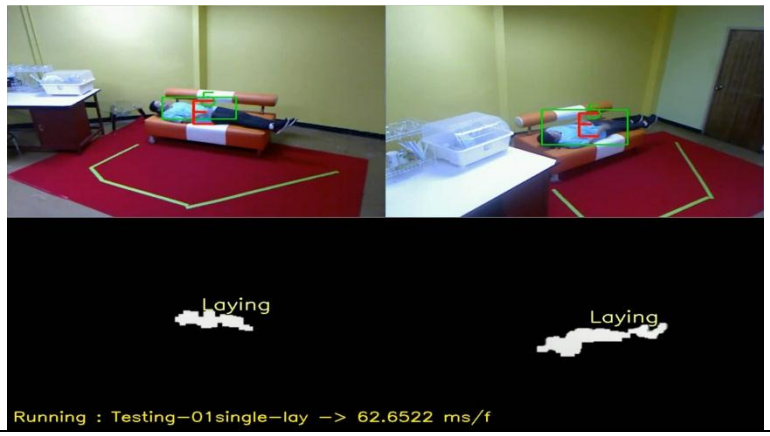
ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก	ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก
#1 A		#4 D	
#2 B		#5 E	
#3 C		-	-

ตารางที่ จ-2 ตัวอย่างการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละหนึ่งคนสำหรับ Dataset # 1

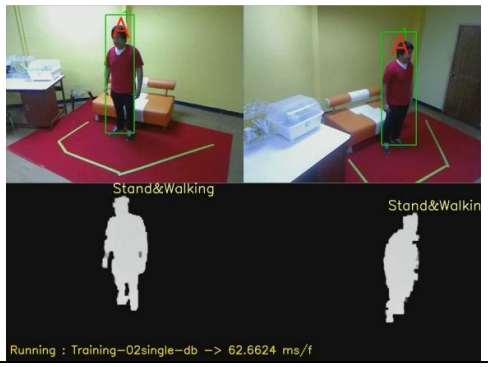
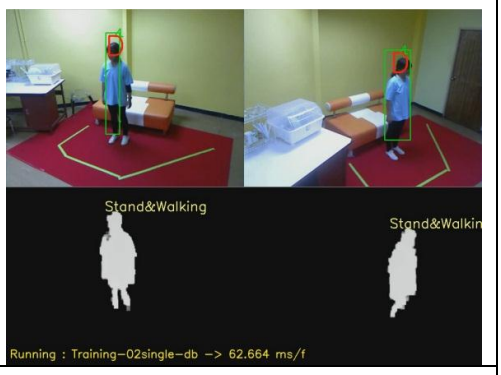
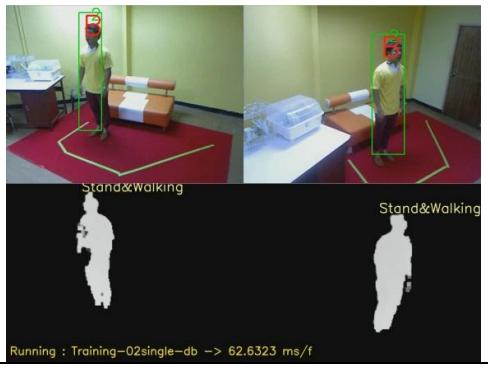
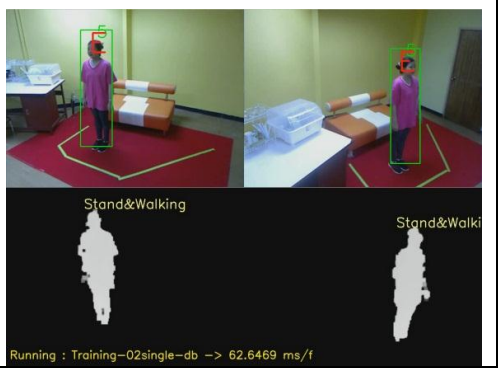
ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
1	
2	
3	

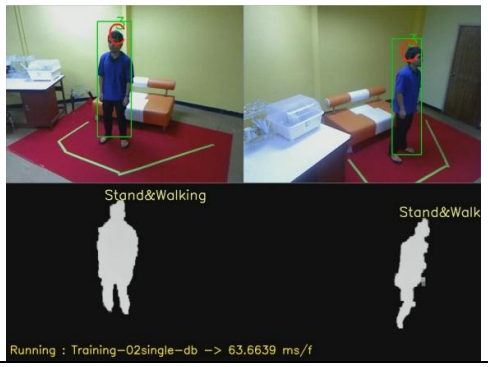
ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
4	
5	
6	

ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
7	
8	
9	

ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
10	

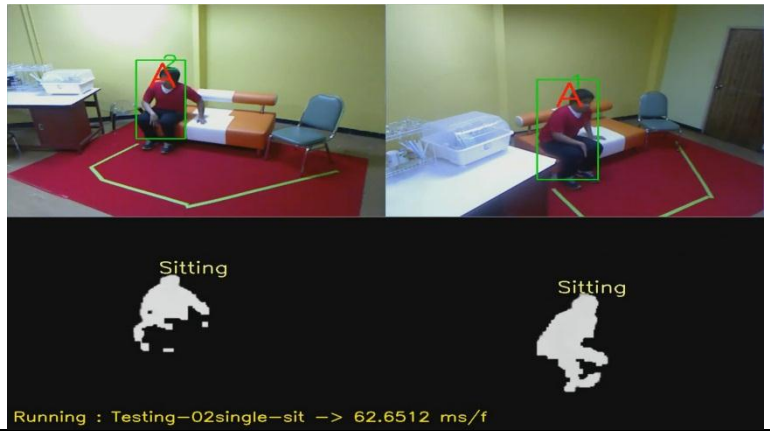
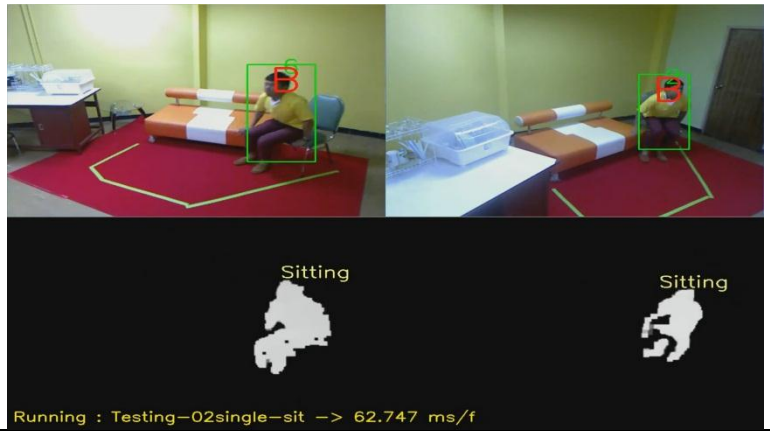
ตารางที่ จ-3 ตัวอย่างบุคคลที่เข้าทดสอบครั้งละหนึ่งคนและ Global ID สำหรับ Dataset # 2

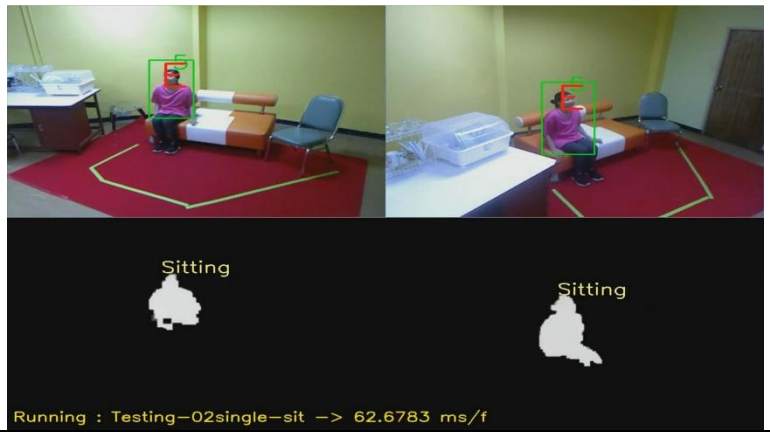
ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก	ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก
# 1 A		# 4 D	
# 2 B		# 5 E	

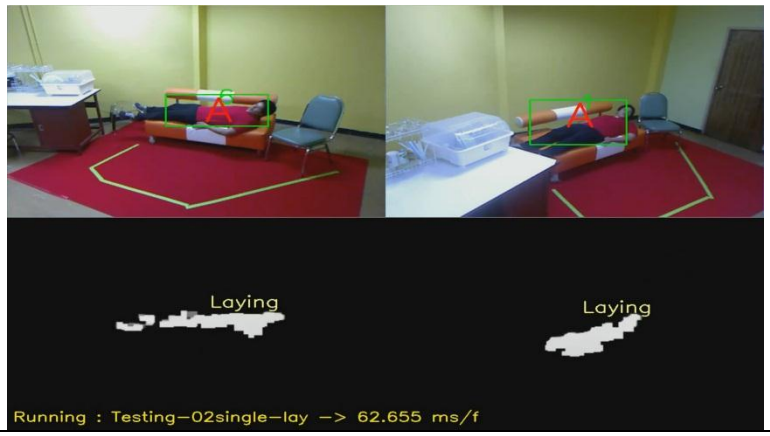
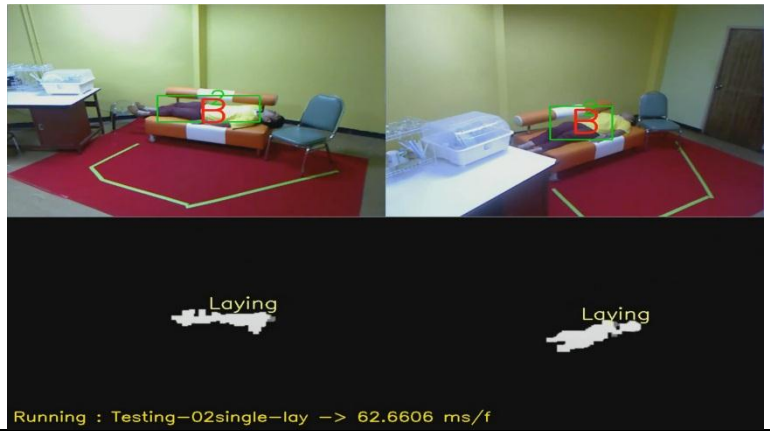
ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก	ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก
# 3 C		-	-

ตารางที่ จ-4 ตัวอย่างการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละหนึ่งคนสำหรับ Dataset # 2

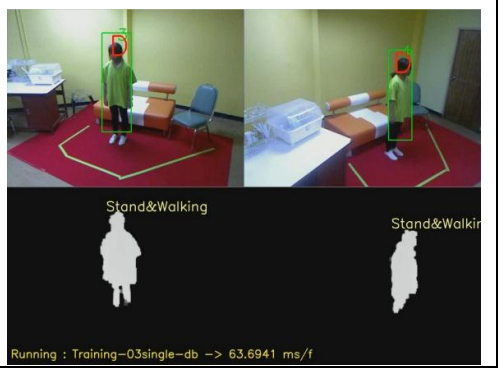
ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
1	
2	

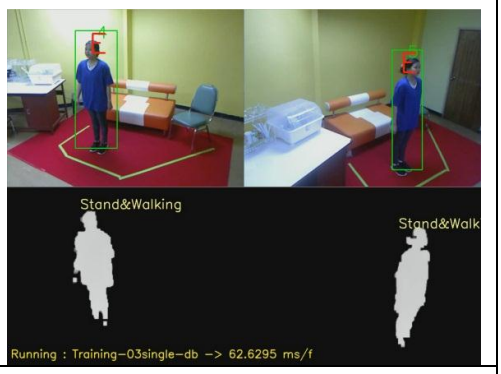
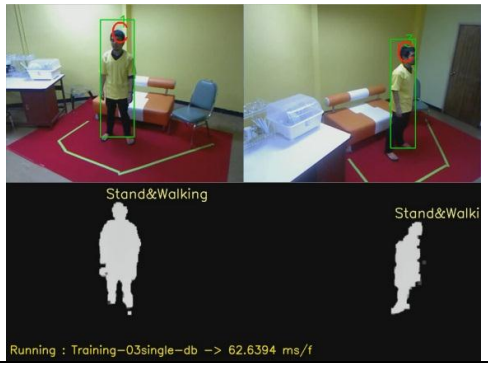
ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
3	
4	
5	

ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
6	
7	
8	

ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
9	
10	

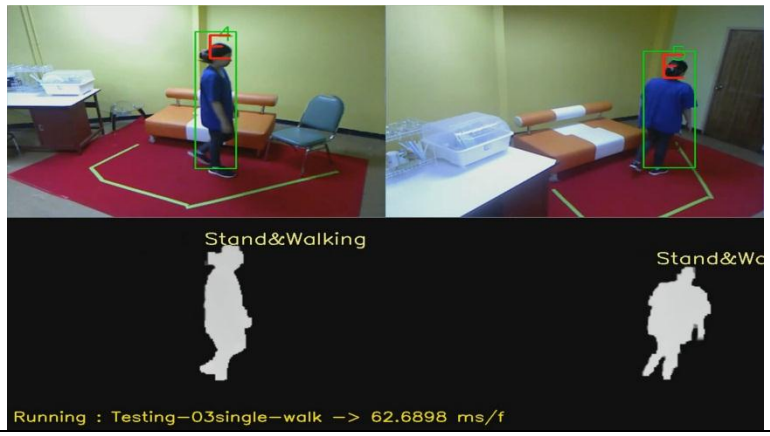
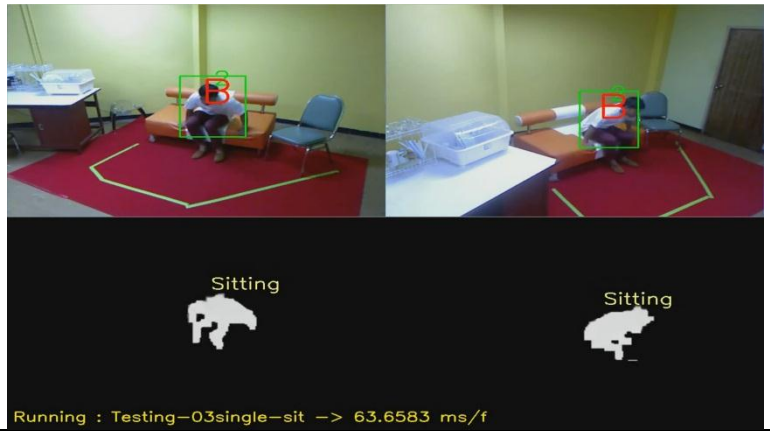
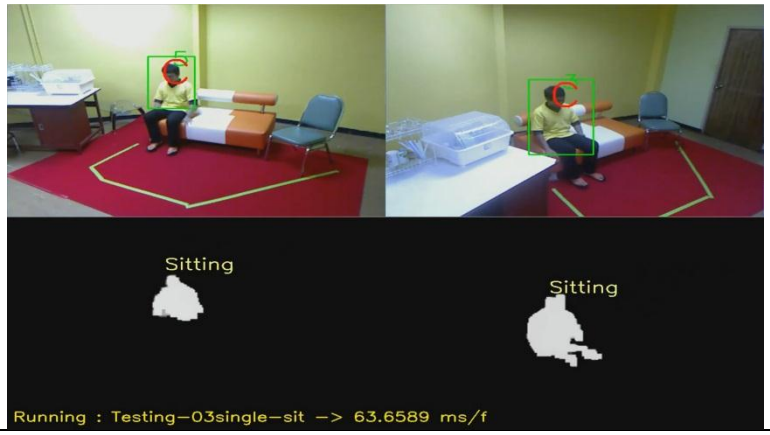
ตารางที่ จ–5 ตัวอย่างบุคคลที่เข้าทดสอบครั้งละหนึ่งคนและ Global ID สำหรับ Dataset # 3

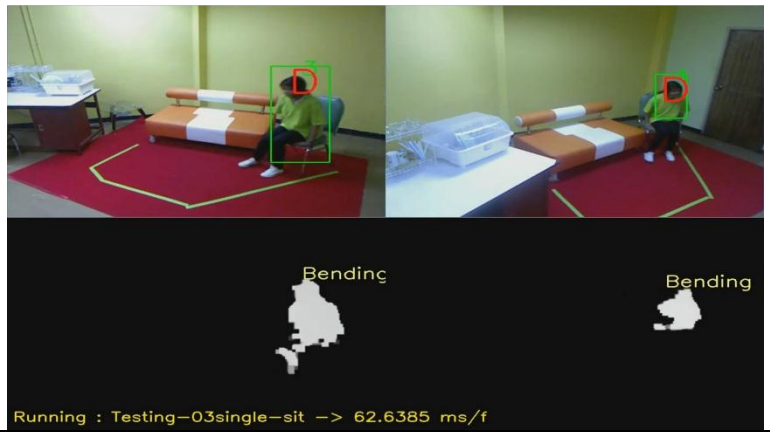
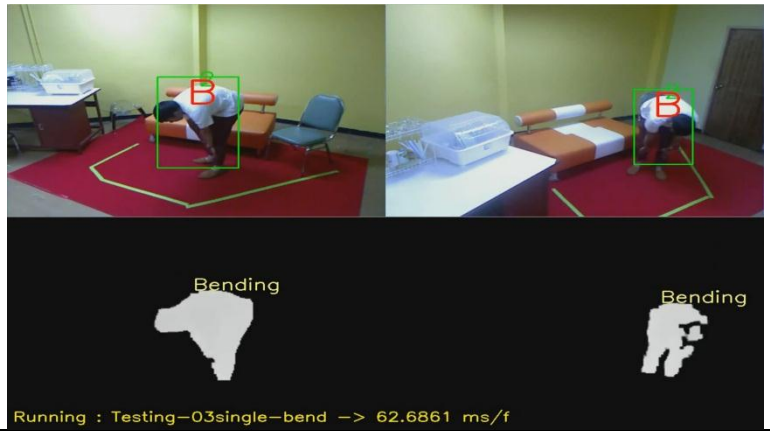
ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก	ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก
#1 A		#4 D	

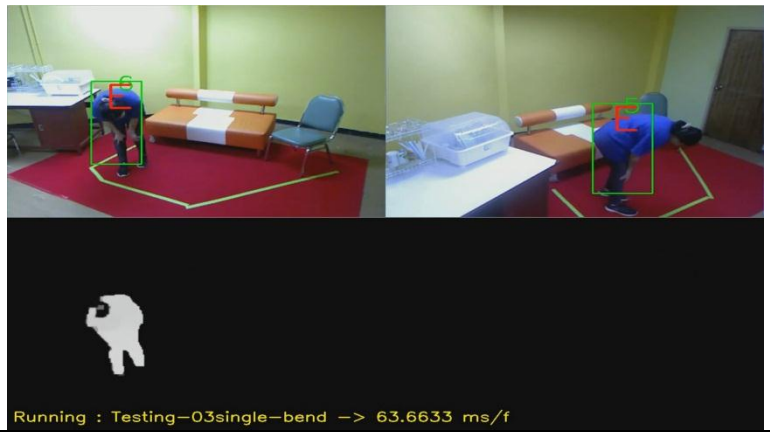
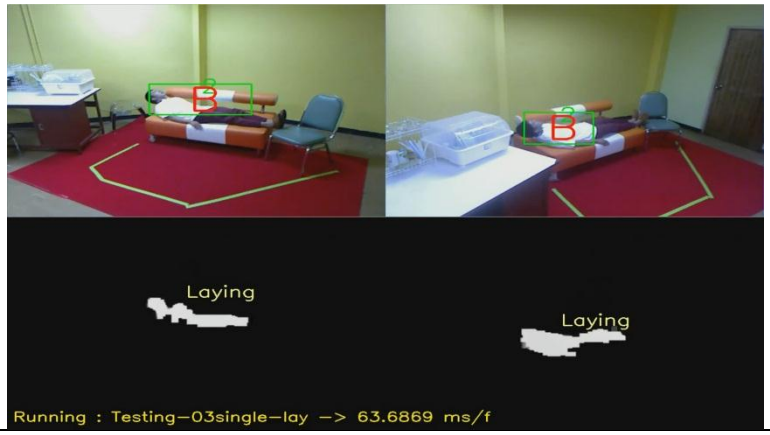
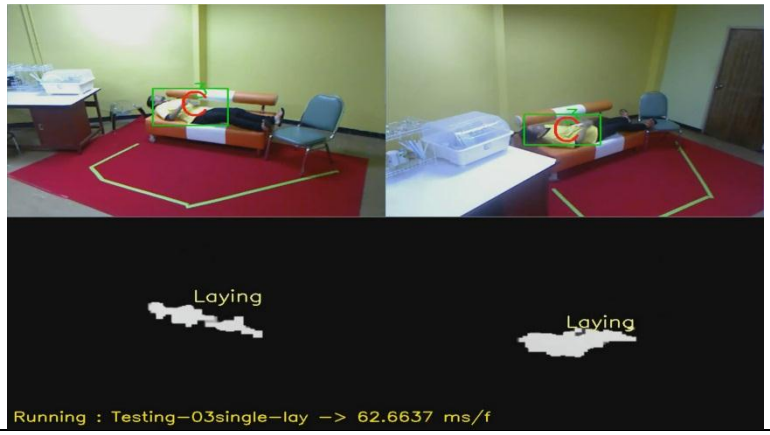
ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก	ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก
#2 B		#5 E	
#3 C		-	-

ตารางที่ จ-6 ตัวอย่างการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละหนึ่งคนสำหรับ Dataset # 3

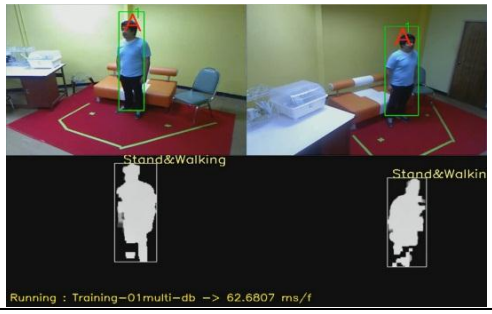
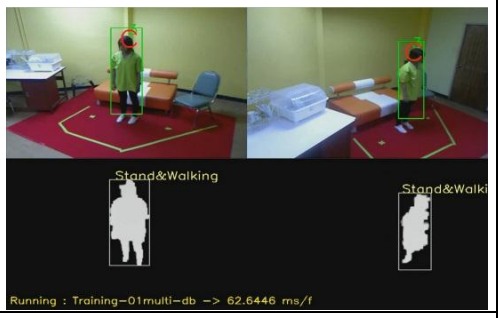
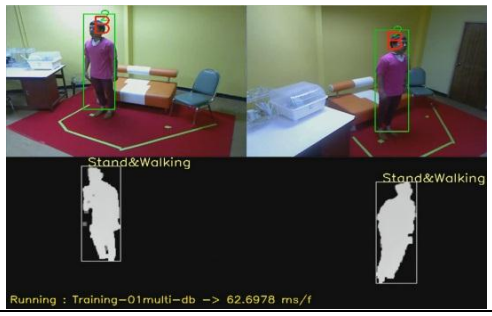
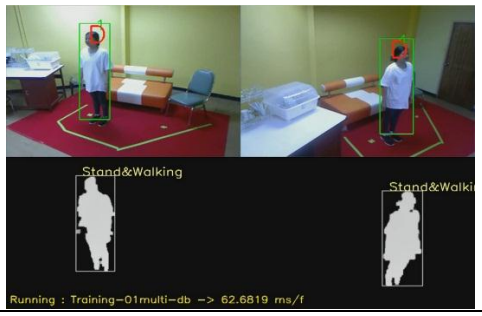
ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
1	

ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
2	
3	
4	

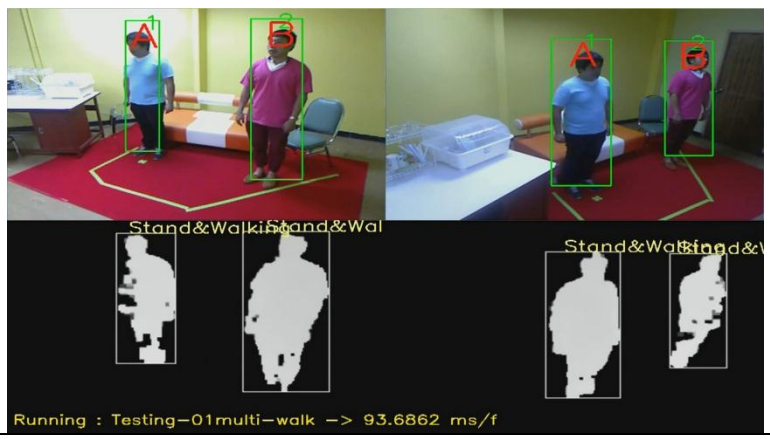
ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
5	
6	
7	

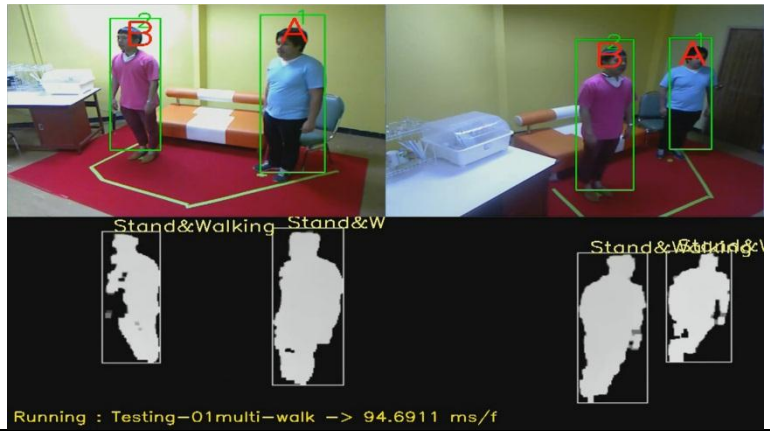
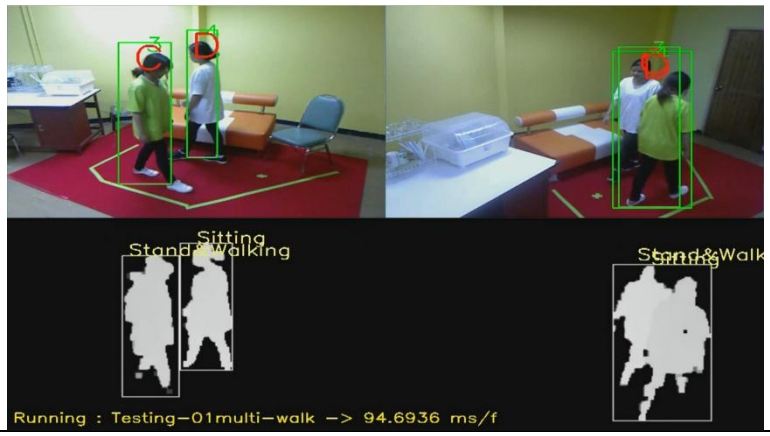
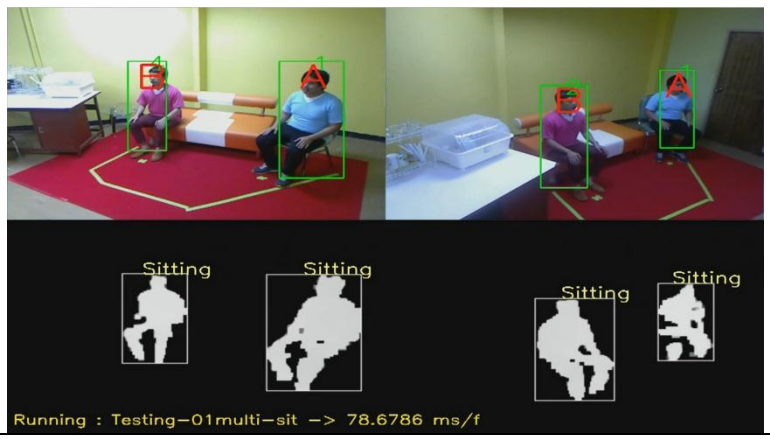
ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
8	
9	
10	

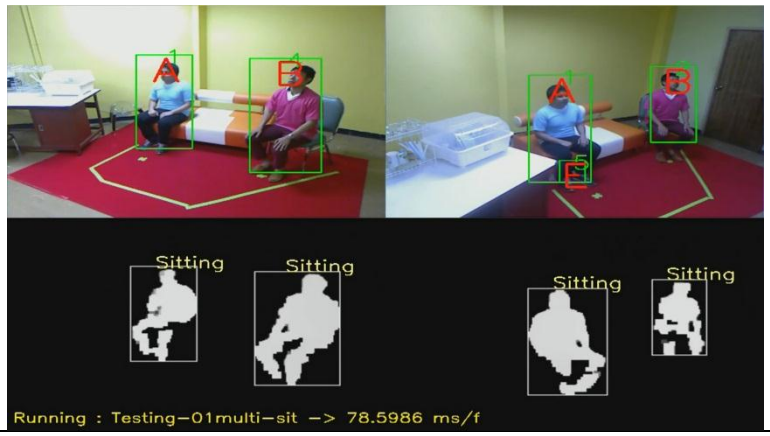
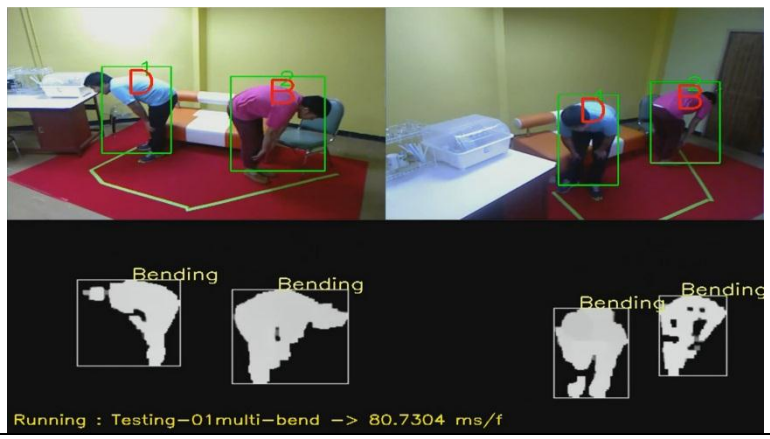
ตารางที่ จ-7 ตัวอย่างบุคคลที่เข้าทดสอบครั้งละสองคนและ Global ID สำหรับ Dataset # 1

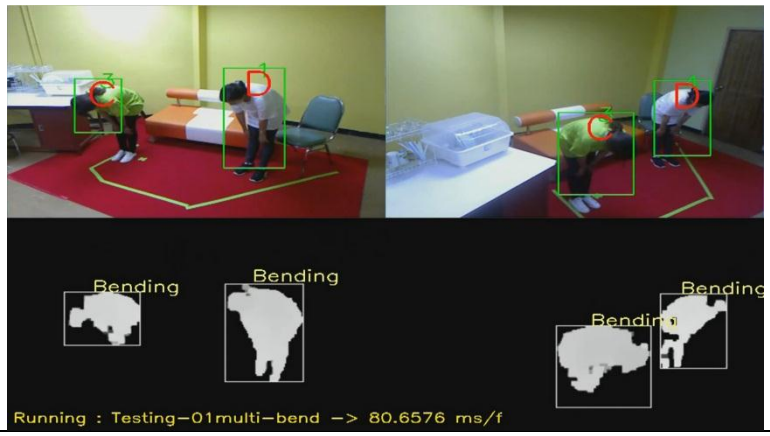
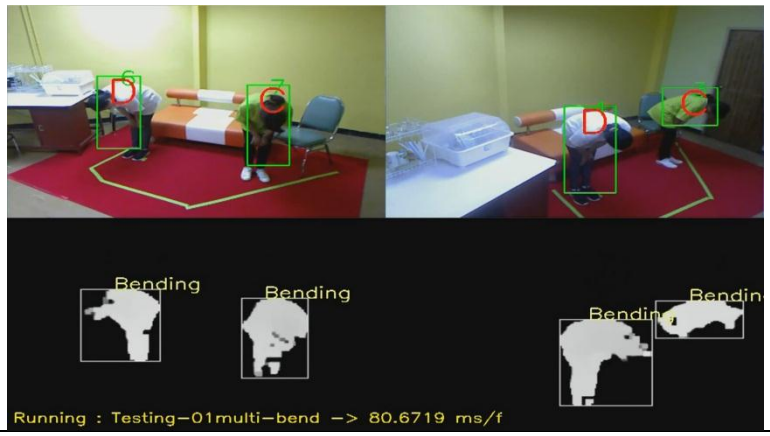
ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก	ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก
# 1 A		# 3 C	
# 2 B		# 4 D	

ตารางที่ จ-8 ตัวอย่างการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละสองคนสำหรับ Dataset # 1

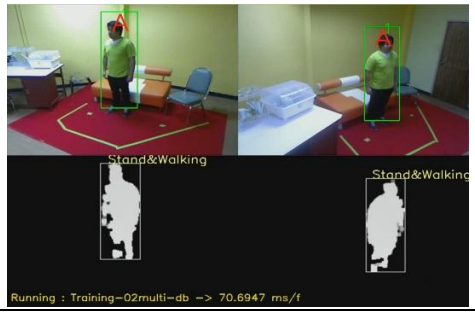
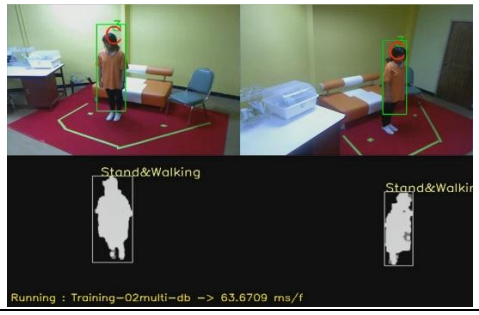
ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
1	

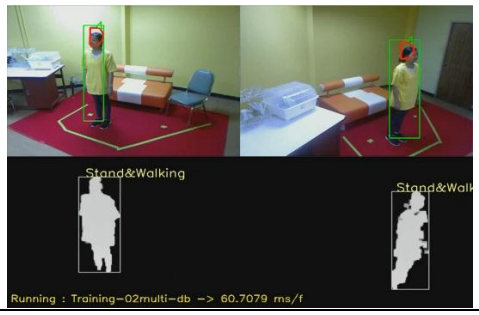
ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
2	
3	
4	

ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
5	
6	
7	

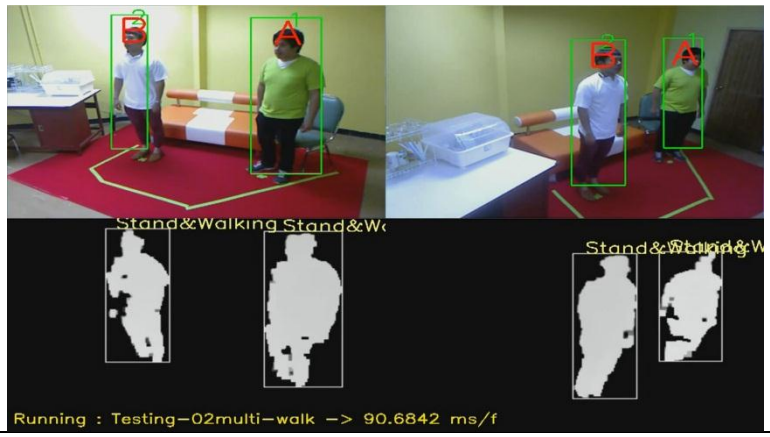
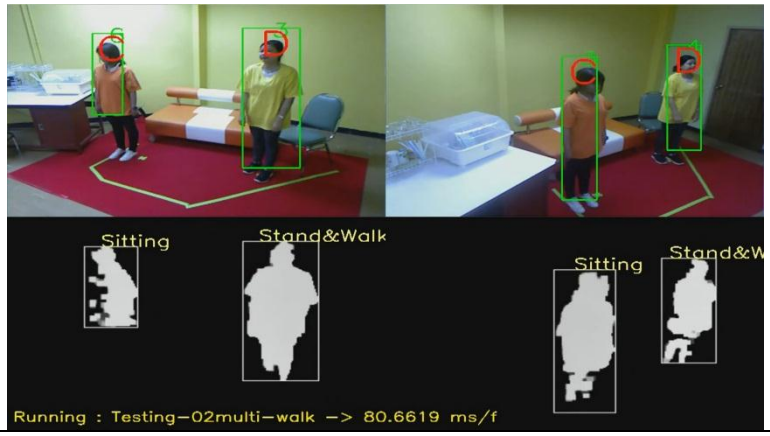
ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
8	
9	

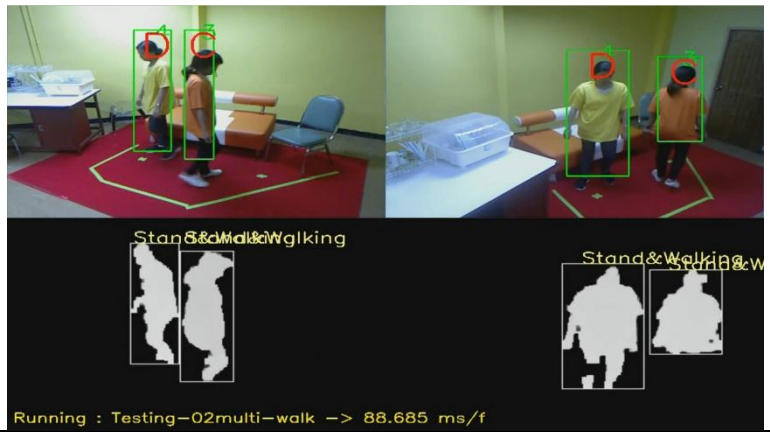
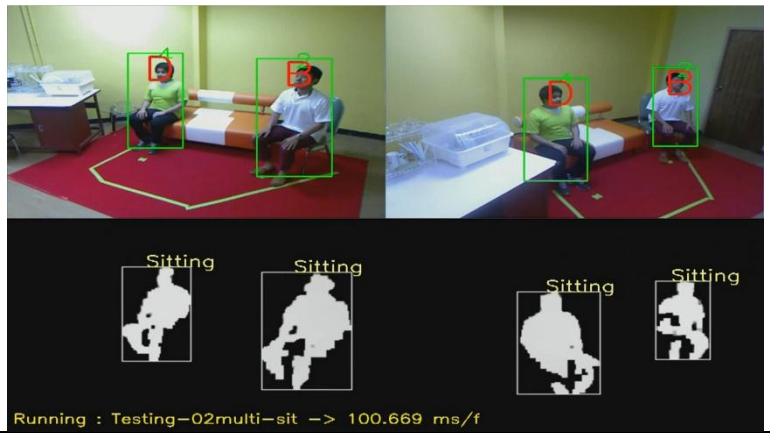
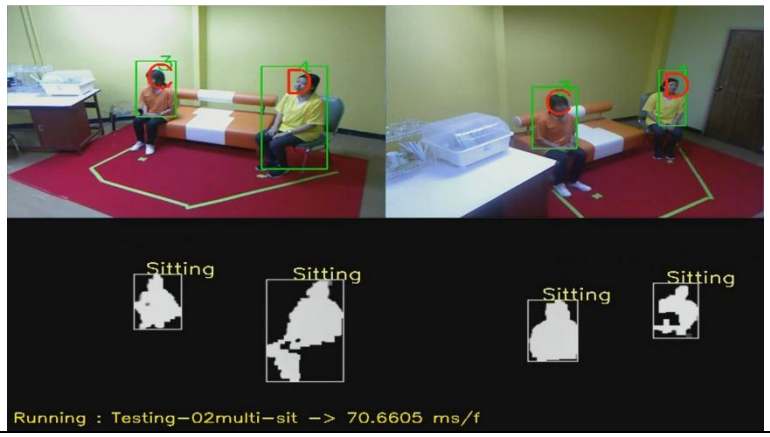
ตารางที่ จ-9 ตัวอย่างบุคคลที่เข้าทดสอบครั้งละสองคนและ Global ID สำหรับ Dataset # 2

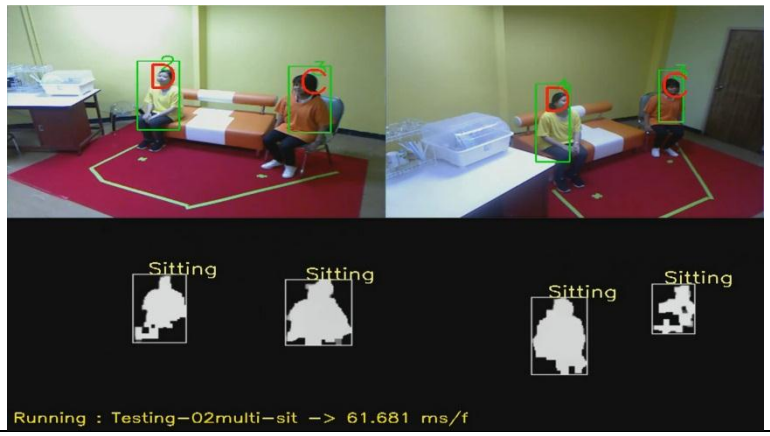
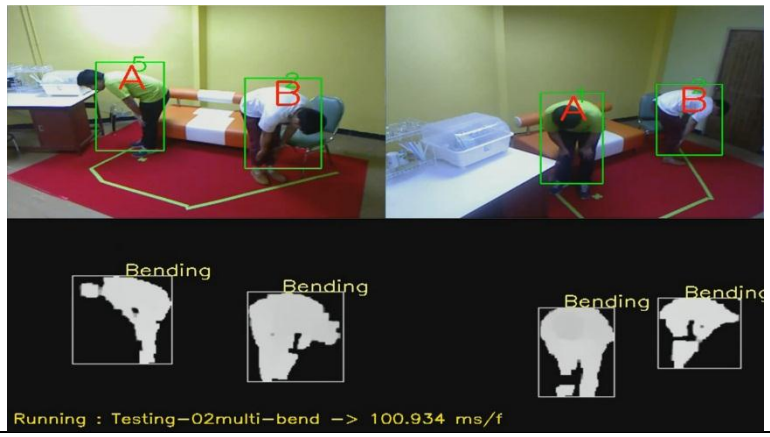
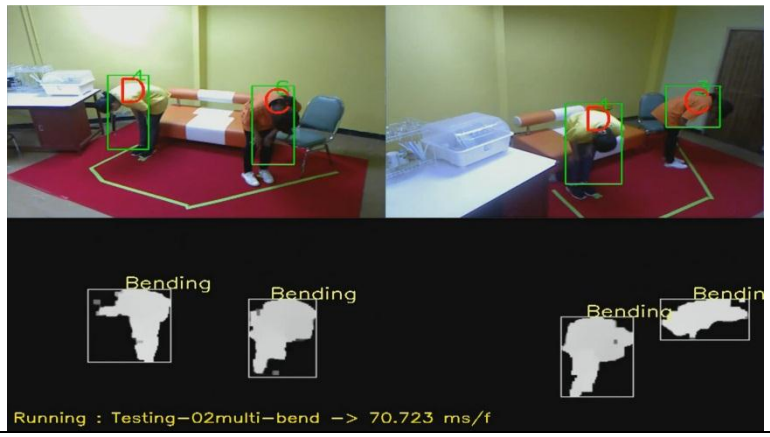
ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก	ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก
#1 A		#3 C	

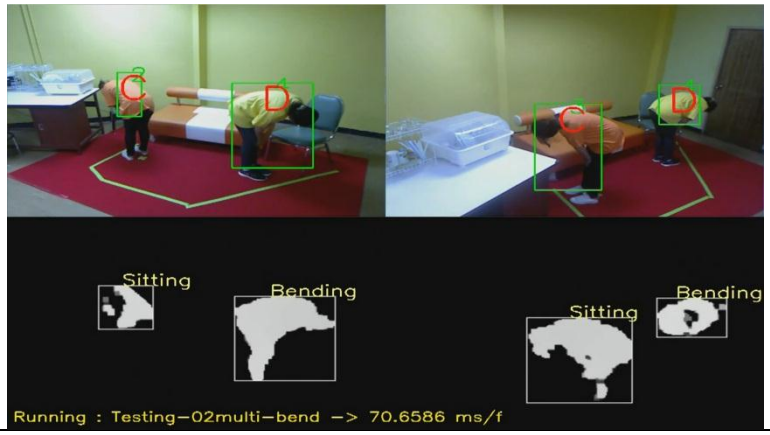
ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก	ID	บุคคลที่เข้าทดสอบและ Global ID ที่ถูก Assign ครั้งแรก
#2 B		#4 D	

ตารางที่ จ-10 ตัวอย่างการทดสอบติดตามและจดจำโดยเข้าไปในระบบครั้งละสองคนสำหรับ Dataset # 2

ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
1	
2	

ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
3	
4	
5	

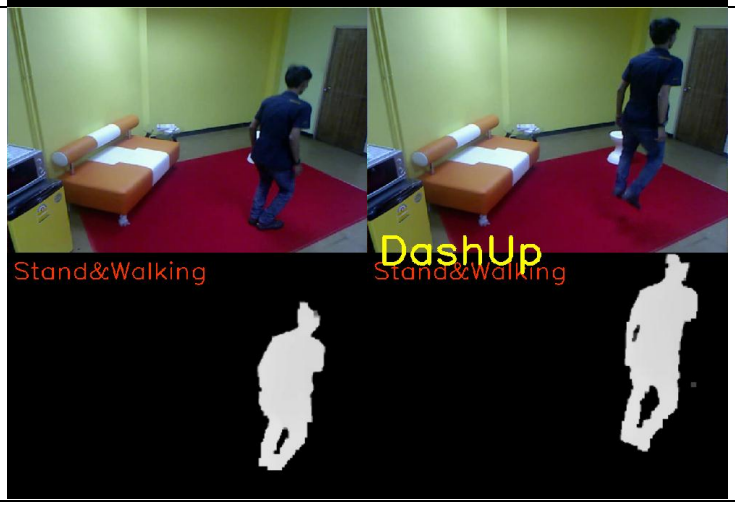
ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
6	
7	
8	

ตัวอย่างที่	ตัวอย่างผลการทดสอบการติดตามและจดจำตัวบุคคลจากท่าทางหลายมุมมอง
9	

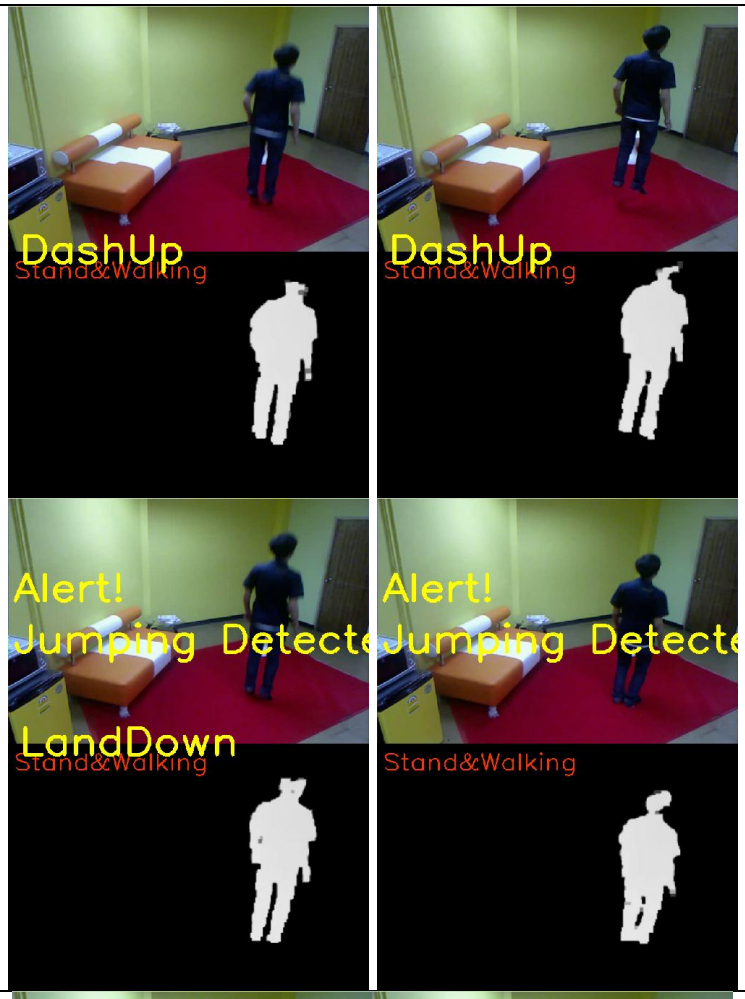
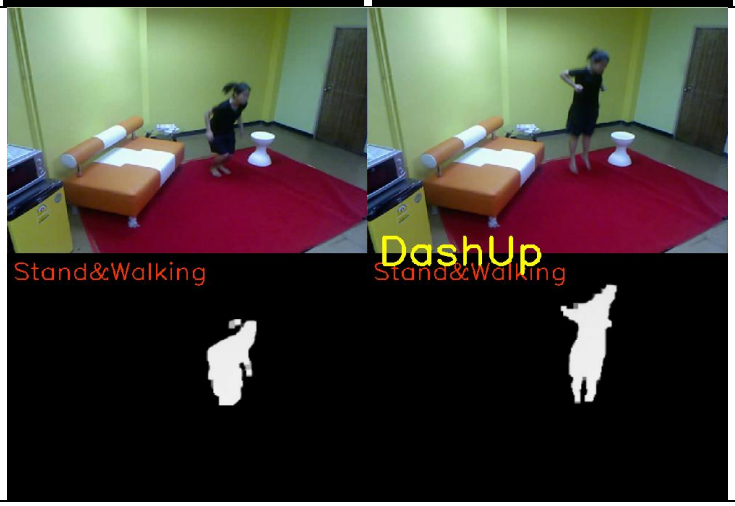
ภาคผนวก ฉ.

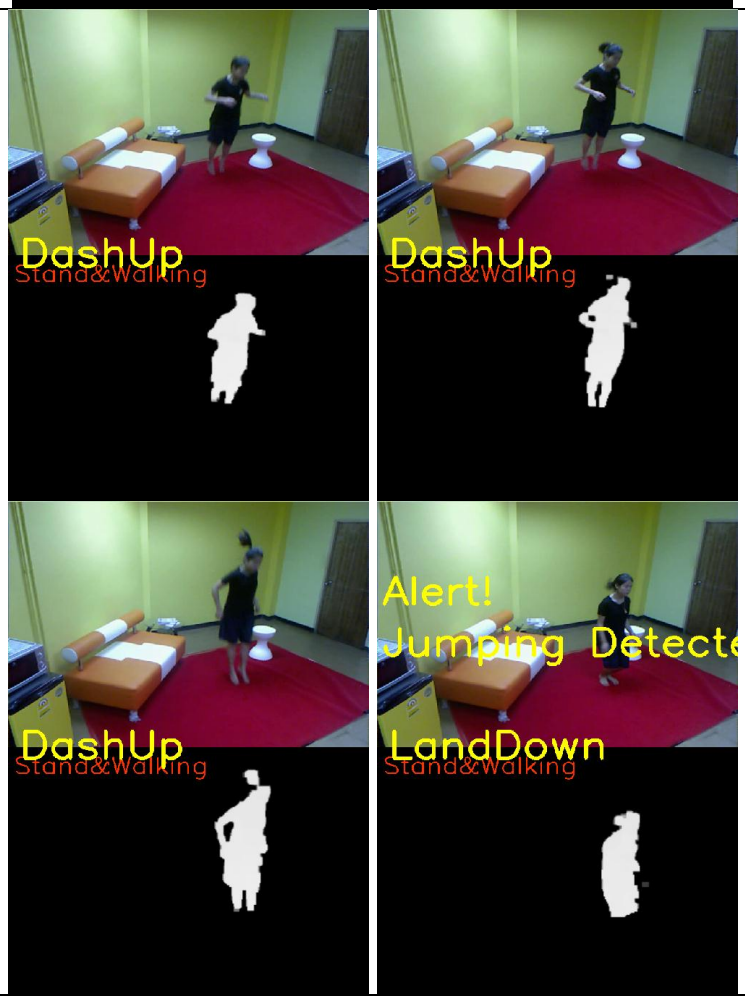
ตัวอย่างผลการทดสอบการตรวจจับท่าทางที่ผิดปกติ:กรณีศึกษาการกระโดด (เพิ่มเติม)

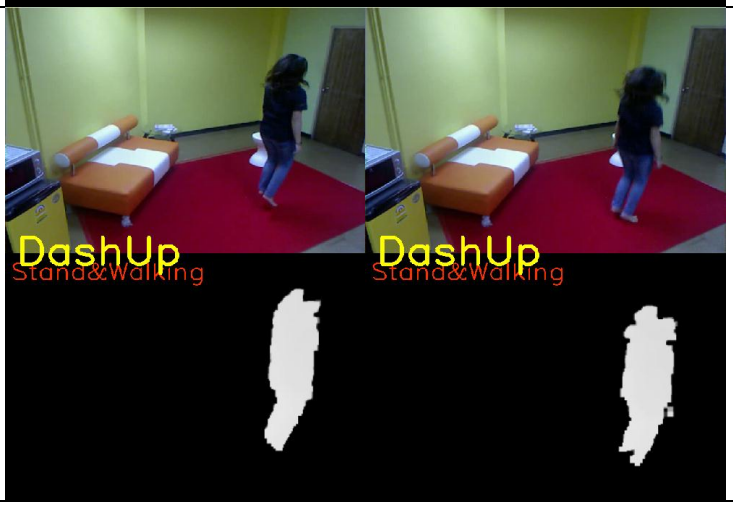
ตารางที่ ฉ-1 ตัวอย่างผลการทดสอบกรณีศึกษาการกระโดด

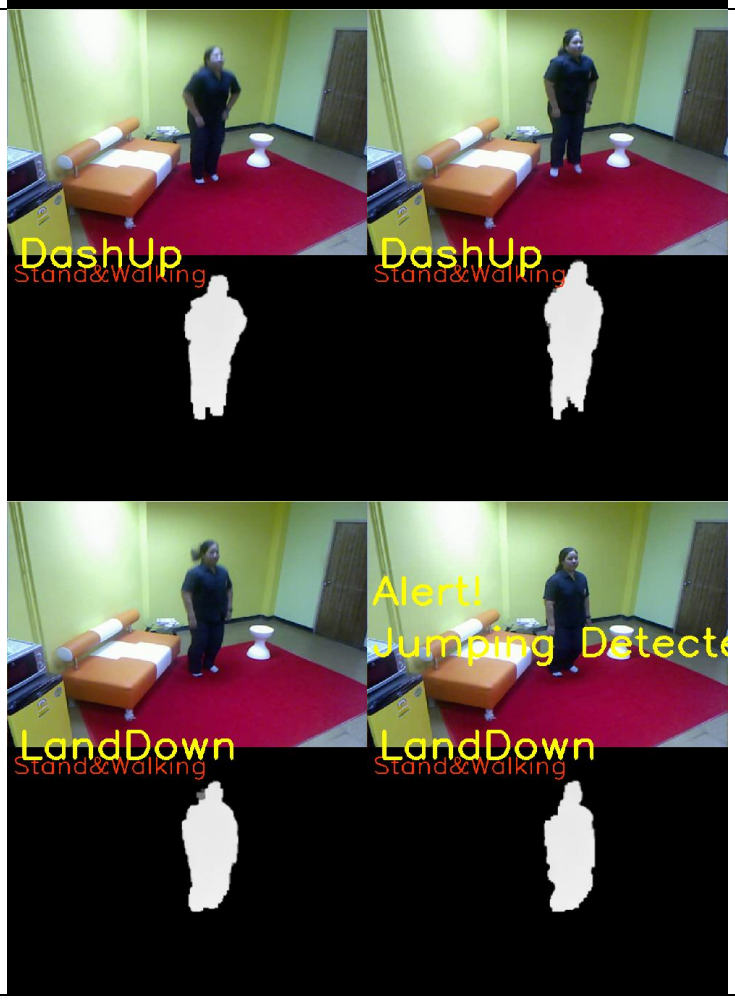
ตัวอย่างที่	ภาพตัวอย่างของการตรวจจับการกระโดด
1	
2	

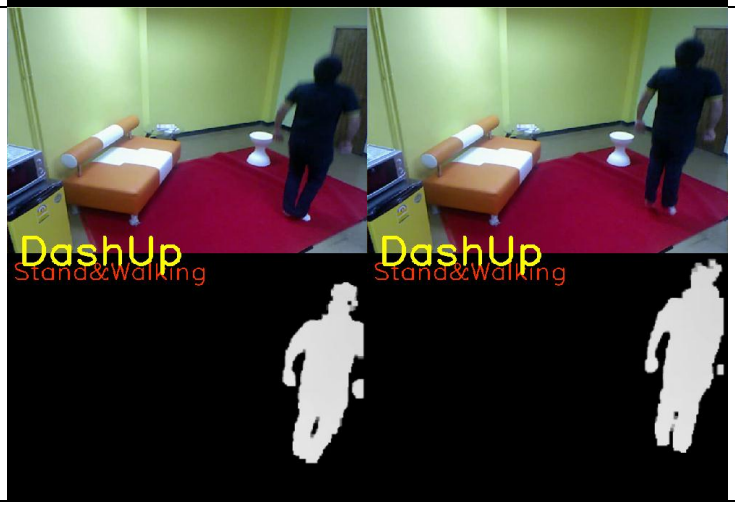
ตัวอย่างที่	ภาพตัวอย่างของการตรวจจับการกระโดด
	
3	

ตัวอย่างที่	ภาพตัวอย่างของการตรวจจับการกระโดด
4	
5	

ตัวอย่างที่	ภาพตัวอย่างของการตรวจจับการกระโดด
	
6	

ตัวอย่างที่	ภาพตัวอย่างของการตรวจจับการกระโดด
7	
8	

ตัวอย่างที่	ภาพตัวอย่างของการตรวจจับการกระโดด
	
9	

ตัวอย่างที่	ภาพตัวอย่างของการตรวจจับการกระโดด
10	
11	

ตัวอย่างที่	ภาพตัวอย่างของการตรวจจับการกระโดด
	
12	 

ประวัติผู้เขียน

ชื่อ สกุล นายพงศกร เจริญเนตรกุล

รหัสประจำตัวนักศึกษา 5610130003

วุฒิการศึกษา

วุฒิ	ชื่อสถาบัน	ปีที่สำเร็จการศึกษา
วิศวกรรมศาสตรบัณฑิต (วิศวกรรมคอมพิวเตอร์)	มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย	2555

ทุนการศึกษา

1. ทุนการศึกษาตามโครงการผลิตและพัฒนาบุคลากรมหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัยสำหรับบุคคลทั่วไป ประจำปีการศึกษา 2556
2. ทุนอุดหนุนการวิจัยเพื่อวิทยานิพนธ์ บัณฑิตวิทยาลัย มหาวิทยาลัยสงขลานครินทร์ ประจำปีงบประมาณ 2558

การตีพิมพ์เผยแพร่ผลงาน

1. Pongsagorn Chalearnnetkul and Nikom Suvonvorn, "High Level Fusion of Profile-Based Human Action Recognition using Multi-view RGBD Information," in *Proceedings of the International Joint Conference on Computer Science and Software Engineering, JCSSE 2015*, 2015, pp. 36-40.
2. Pongsagorn Chalearnnetkul and Nikom Suvonvorn, "A Rectangular Layer Model for Profile-Based Human Action Recognition using Multi-view Depth Information," in *Asia-Pacific Journal of Science and Technology(APST)*, Vol. 22, No. 3, Sep 30, 2017.
3. Pongsagorn Chalearnnetkul and Nikom Suvonvorn, "Multiview Layer Fusion Model for Action Recognition Using RGBD Images," *Computational Intelligence and Neuroscience*, vol. 2018, Article ID 9032945, 22 pages, 2018. https://doi.org/10.1155/2018/9032945.