JIST Journal of Information Science and Technology
Volume 12, NO 1 | JAN – JUN 2022 | 27-36
ISSN: 2651-1053 (Online)
ISSN: 1906-9553 (Print)

แบบจำลองการเรียนรู้ท่าทางมนุษย์จากการเคลื่อนไหวด้วยเทคนิคโครงข่ายประสาทเทียมแบบสังวัตนาการและแบบหน่วยความจำระยะสั้นแบบยาว

Learning Model of Human Body Movement using Convolutional Neural Network and Long Short-Term Memory

นัศพ์ชาณัณ ชินปัญช์ธนะ

Nutchanun Chinpanthana

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

College of Innovative Technology and Engineering, Dhurakij Pundit University

Received: December 17, 2021; Revised: May 03, 2022; Accepted: May 18, 2022; Published: June 29, 2022

ABSTRACT – The recognition and understanding of human activities remain active research in computer vision and it is widely used in computer interaction, human monitoring system, human behavior analysis, and other fields. The problem is challenging due to its various interactions between persons and motion dynamics between human and objects. There are many algorithms to recognize human activity based on deep neural network models was proposed. The deep neural network classifier was constructed to integrate with the human action target, but the results was not classified as action though it should have. To overcome this problem, this paper proposes a combined model of deep learning convolutional network with long short-term memory for recognize human motion pose. The proposed model extracts the key point features in an automated way with OpenPose and categorizes with trajectories of a human body. This paper presents main three steps including (1) data pre-processing and feature extraction, (2) deep learning with convolutional neural network and long short-term memory, and (3) efficiency measurement and evaluation of experimental results. Results we have tested our approach on activities of daily living with URFall datasets. The experimental results indicate that our framework offers performance improvements. The proposed model can achieve significant improvements for activity classification with maximum success rate of 81% and 76.9% with 256, 128 hidden nodes in data set II.

KEYWORDS: Image classification, Human recognition, Convolutional Neural Network, Human activity recognition

บทคัดย่อ -- งานวิจัยด้านคอมพิวเตอร์วิชันในส่วนของการรู้จำและเข้าใจกิจกรรมมนุษย์ยังคงวิจัยที่สำคัญสำหรับการมีปฏิสัมพันธ์ระหว่างมนุษย์ ไม่ว่าจะเป็นระบบเฝ้าระวัง การวิเคราะห์พฤติกรรมมนุษย์ เป็นต้น ปัญหาที่เกิดขึ้นเมื่อมีการปฏิสัมพันธ์ระหว่างบุคคลหลายคน การเคลื่อนไหวระหว่างบุคคลและวัตถุที่ซับซ้อน มีหลายอัลกอริธึมที่พยายามรวมคุณลักษณะต่างๆเพื่อนำมาใช้ในการวิเคราะห์การรู้จำท่าทางมนุษย์ด้วยรูปแบบโครงข่ายประสาทเทียมเชิงลึก เทคนิคนี้ถูกสร้างขึ้นสำหรับการจำแนกด้วยท่าทางมนุษย์โดยเฉพาะแต่ผลที่ได้ยังไม่ได้ตามที่กำหนดไว้ ทำให้ยังคงเป็นปัญหา ดังนั้นในงานวิจัยนี้จึงได้นำเสนอการเรียนรู้แบบจำลองที่มีการผสมผสานการเรียนรู้ด้วยเทคนิคโครงข่ายประสาทเทียมแบบสังวัตนาการและแบบหน่วยความจำระยะสั้นแบบยาว การสกัดคุณลักษณะตำแหน่งสำคัญด้วย OpenPose เพื่อนำมาใช้ในการจำแนกร่วมกับข้อมูลการเคลื่อนที่ของร่างกาย วิธีการดำเนินงานประกอบด้วย 3 ขั้นตอนหลักดังนี้ (1) การเตรียมข้อมูล และสกัด

Corresponding Author: nutchanun.cha@dpu.ac.th

ข้อมูล (2) การเรียนรู้เชิงลึกด้วยเทคนิคโครงข่ายประสาทเทียมแบบสังวัตนาการและแบบหน่วยความจำระยะสั้นแบบยาว และ (3) การวัดประสิทธิภาพและประเมินผลจากการทดลอง ผลลัพธ์จากการทดลองด้วยกิจกรรมในชีวิตประจำวันในฐานข้อมูล URFall สามารถระบุได้ว่าการใช้แบบจำลองการเรียนรู้เชิงลึกด้วยเทคนิคโครงข่ายประสาทเทียมแบบสังวัตนาการและแบบหน่วยความจำระยะสั้นแบบยาวร่วมกับคุณลักษณะการเคลื่อนที่ของมนุษย์สามารถเพิ่มประสิทธิภาพของการจำแนกท่าทางได้ดียิ่งขึ้นถึงร้อยละ 81 และ ร้อยละ 76.5 ด้วยโหนดซ่อน 256 และ 123 ในชุดข้อมูลที่ 2

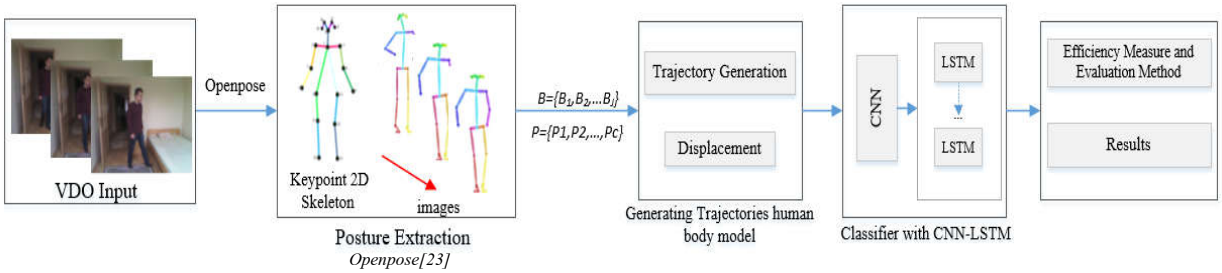
คำสำคัญ: การจำแนกข้อมูลภาพ, การรู้จำมนุษย์, โครงข่ายประสาทเทียมแบบสังวัตนาการ, การรู้จำกิจกรรมมนุษย์

1. บทนำ

การวิวัฒนาการของเทคโนโลยีของเทคโนโลยีที่เปลี่ยนแปลงไปอย่างรวดเร็วมีผลทำให้ เครื่องมืออุปกรณ์อิเลคทรอนิกส์ รวมทั้งอุปกรณ์ถ่ายภาพดิจิทัลถูกพัฒนาอย่างต่อเนื่อง ทำให้มีงานวิจัยในสาขาคอมพิวเตอร์วิชันเกิดขึ้นมากมาย การสร้างแอพพลิเคชันใหม่ๆ [1][2] เพื่อตอบสนองความต้องการของผู้ใช้งานที่หลากหลายตลอดเวลา นักวิจัย Argyle M. [3] ได้ใช้คุณลักษณะข้อมูลพื้นฐานของภาพที่ถูกสกัดขึ้นมาเพื่อจะแปลงข้อมูลเป็นท่าทางของมนุษย์ เช่น นั่ง ยืน เดิน และนอน โดยใช้ความสัมพันธ์กับข้อมูลตำแหน่งของร่างกาย เช่น ศีรษะ แขน ขา ลำตัว และ มุมองศา รวมทั้งการเอนเอียงของร่างกาย ในปัจจุบันการวิเคราะห์ท่าทางมนุษย์ กลายเป็นสิ่งสำคัญที่ถูกนำมาประยุกต์ใช้ ในการวิเคราะห์ให้สามารถตีความหมายข้อมูลบนภาพ เช่น การตรวจจับการเคลื่อนไหวของมนุษย์เพื่อรักษาความปลอดภัย [4] การวิเคราะห์โรคด้วยการเคลื่อนไหวท่าทางมนุษย์เพื่อวิเคราะห์การรักษาผู้ป่วย [5]-[7] การสื่อสารด้วยสัญลักษณ์จากท่าทางการเคลื่อนไหวของมนุษย์สำหรับผู้พิการทางการได้ยิน [8] การโต้ตอบบนเกมส์บนแอพพลิเคชันต่างๆ เป็นต้น แต่อย่างไรก็ตามกระบวนการวิเคราะห์ดังกล่าวนั้น มีความแปรเปลี่ยนไปตามสภาพแวดล้อมและรูปแบบของการสกัดข้อมูล และยังไม่มีการสร้างเป็นแบบจำลองมาตราฐานสำหรับท่าทางจากการเคลื่อนไหวของมนุษย์ในชีวิตประจำวัน เพื่อเป็นมาตรฐานสำหรับรูปแบบการเคลื่อนไหวด้วยแนวการเคลื่อนที่จากจุดสำคัญในตำแหน่งข้อต่อมนุษย์ ดังนั้นในงานวิจัยนี้จึงได้นำเสนอการสร้างแบบจำลองการเรียนรู้ท่าทางมนุษย์จากการเคลื่อนไหวด้วย เทคนิคโครงข่ายประสาทเทียมแบบสังวัตนาการร่วมกับ แบบหน่วยความจำระยะสั้นแบบยาวเพื่อทำให้สามารถนำมาใช้ในการวิเคราะห์ท่าทางร่วมกับงานอื่นๆได้อย่างมีประสิทธิภาพ

2. งานวิจัยที่เกี่ยวข้อง

การรู้จำท่าทางที่เกิดขึ้นในชีวิตประจำวันเป็นพื้นฐานของการเข้าใจถึงกระทำกิจกรรมต่างๆของมนุษย์ [9] เช่น การเดิน, การวิ่ง,การรับประทานอาหาร, การรับโทรศัพท์ เป็นต้น สำหรับการเคลื่อนไหวของร่างกายเพื่อทำกิจกรรมใดกิจกรรมหนึ่ง จะถูกวิเคราะห์จากความแตกต่างกันของลำดับตำแหน่งของร่างกาย รวมทั้งความเร็วที่ร่างกายเคลื่อนไหว Wang et al., [10][11] ได้วิเคราะห์การกระทำของมนุษย์ที่แสดงออกบนภาพ ด้วยวิธีการหาความแตกต่างของคัสเตอร์ท่าทางมนุษย์จากการจับคู่ด้วยการแปลงรูปทรงวัตถุ (Deformable shape matching) งานวิจัยของ Thurnau และ Hlavac [12] ใช้การแทนรูปแบบของท่าทางด้วยฮีสโตแกรมของท่าทางหลักโดยใช้ Nonnegative matrix factorization และกลุ่ม A. Andre Chaaraoui et al., [13] ได้พยายามสกัดคุณลักษณะ จากท่าทางมนุษย์จากภาพเงา (Silhouettes) แบบ 1 มิติที่มีการแทนค่าเป็นโปรเจคชันของร่างกายมนุษย์ แขนและขารวม 6 ส่วนแต่การสกัดฟีเจอร์ด้วยภาพเงาในบางครั้งการทับกันของส่วนร่างกายอาจจะทำให้การสกัดฟีเจอร์ได้ไม่ครบ แต่อย่างไรก็ตาม S. Yi [14] ได้พยายามจำแนกท่าทางของมนุษย์ ด้วยวิธี Sparse Granger Causality Graph Model เป็นการแทนตำแหน่งและความสัมพันธ์ของร่างกายด้วยโหนดและเส้นเชื่อมโยงบนกราฟ ร่างกายมนุษย์ถูกสกัดด้วยวิธี canny operator เป็นขอบของโครงร่างกาย การเคลื่อนไหวของร่างกายดูจากค่าความเข้มของสีที่เปลี่ยนแปลงไป และการเปรียบเทียบกันของจำนวนพิกเซลบนภาพดิจิทัลเพื่อใช้บอกระยะห่างและใช้โมเดลการจับคู่ เพื่อบอกตำแหน่งร่างกายของมนุษย์ มีนักวิจัยบางกลุ่มพยายามใช้เซ็นเซอร์ Smart wearable อุปกรณ์พกพาในการตรวจจับกิจกรรมของมนุษย์ เพื่อแยกแยะผู้สูงอายุที่กำลังจะล้ม งานวิจัยกลุ่ม Alzahrani et al. [18] ได้พยายามปรับปรุงการรู้จำท่าทางการเคลื่อนไหวของมนุษย์



รูปที่ 1. ขั้นตอนการเรียนรู้ท่าทางมนุษย์จากการเคลื่อนไหว

ด้วยการแทนโครงสร้างมนุษย์ด้วยโครงร่างกระดูก (skeleton) เพื่อวิเคราะห์ข้อมูลแบบระบบเวลาจริง ด้วย โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ (Backpropagation neural network) และ มีนักวิจัยกลุ่ม Ghazal et al. [19] Franco et al. [20] เริ่มใช้กล้อง Kinect เพื่อเก็บข้อมูลในรูปแบบ 3 มิติและพยายามสกัดข้อมูลข้อต่อของมนุษย์เพื่อรู้จำท่าทาง และจำแนกออกเป็นท่าทางพื้นฐาน เช่น นั่ง และ ยืน เป็นต้น ยังมีงานวิจัยที่ได้การตรวจจับวัตถุและการติดตามวัตถุที่มีมากกว่าหนึ่งจุดสนใจด้วยการใช้ลำดับของโครงร่างกระดูก [22][23] เป็นตัวแยกแยะคุณลักษณะสำคัญ เพื่อหาความสัมพันธ์ระหว่างโหนดและจุดที่ใกล้เคียงกันระหว่างเฟรม และความต่อเนื่องกันเพื่อสร้างความสัมพันธ์ที่เกิดขึ้นร่วมกันระหว่างโดเมนเชิงพื้นที่และข้อมูลเวลา แต่อย่างไรก็ตามยังเกิดการรบกวนจากพื้นที่โดยรอบและความหมายที่ได้มีมากจนเกินไปการนำกระบวนการเรียนรู้เข้ามาช่วยเพื่อให้ผลได้เป็นพฤติกรรมมนุษย์ จะเห็นว่ากลุ่มนักวิจัยส่วนใหญ่นำเทคนิคเข้ามาประยุกต์ใช้ผ่านรูปแบบการเรียนรู้ตามกิจกรรมของมนุษย์ที่สนใจ ผลที่ได้สามารถใช้ในการจำแนกท่าทางได้บางส่วน ยังมีข้อจำกัดทั้งในรูปของการสกัดข้อมูลจากภาพ เพื่อปรับปรุงการจำแนกท่าทางมนุษย์จากการเคลื่อนไหวให้มีประสิทธิภาพมากขึ้น ดังนั้นในงานวิจัยนี้นำเสนอการจำแนกท่าทางมนุษย์ในชีวิตประจำวันด้วยการสร้างแนวการเคลื่อนที่ จากท่าทางการเคลื่อนไหวด้วยแบบจำลองการเรียนรู้ด้วยเทคนิคโครงข่าย ประสาทเทียมแบบสังวัตนาการร่วมกับเทคนิคแบบหน่วยความจำระยะสั้นแบบยาว (Convolutional Neural Network and Long Short-Term Memory) เพื่อจำแนกท่าทางการเคลื่อนไหวของร่างกายมนุษย์ในกิจวัตรประจำวัน

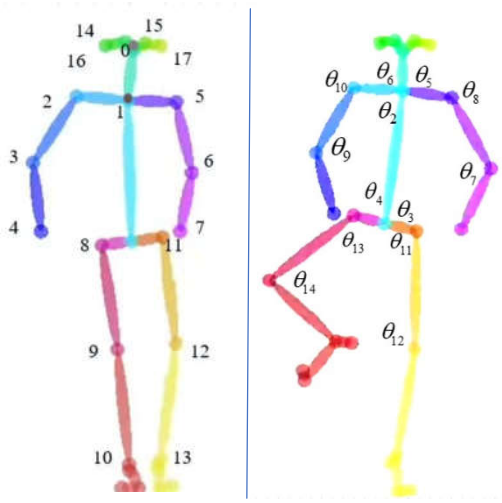
3. วิธีการวิจัย

งานวิจัยนี้ได้นำเสนอการจำแนกท่าทางมนุษย์ ด้วยการใช้พลังงานการเคลื่อนไหวของร่างกายมนุษย์ เพื่อนำมาวิเคราะห์ภาพรวมที่ได้จากท่าทางและความเร็วของการเคลื่อนไหว เป็นทฤษฎีพื้นฐานทางชีววิทยามนุษย์ทางด้านกลศาสตร์ ด้วยการสกัดข้อมูลและตำแหน่งจากการเคลื่อนที่ของร่างกาย (Human action movement) จากแต่ละตำแหน่งบนจุดสำคัญบนร่างกายด้วยมุมองศาที่เกิดขึ้น (Movement trajectory) บนโครงร่างจำลองรูปแบบสองมิติที่เรียกว่า โครงร่างกระดูก แสดงถึงตำแหน่งข้อต่อที่เป็นจุดสนใจทั้งหมด 15 จุด เป็นตำแหน่งอ้างอิง (reference point) ส่วนร่างกายท่อนบน (upper body) และส่วนร่ายกายด้านล่าง (lower body) ประกอบด้วย ส่วนหัว (head), คอ (neck), ไหล่ (shoulder), แขนส่วนบน (upper arms), แขนส่วนล่าง (forearms), สะโพก (hip), เข่า (knee) และ ข้อเท้า (ankle) ทั้งด้านซ้ายและขวา มีขั้นตอนการดำเนินงานวิจัยดังนี้ (1) การเตรียมข้อมูล และสกัดข้อมูล (2) การเรียนรู้เชิงลึกด้วยเทคนิคโครงข่ายประสาทเทียมและ (3) การวัดประสิทธิภาพและการประเมินผลการทดลอง ดังแสดงในรูปที่ 1

3.1 การเตรียมข้อมูล และสกัดข้อมูล

การเตรียมข้อมูล เริ่มจากคัดเลือกฐานข้อมูลภาพวิดีโอที่มีมนุษย์เด่นชัดจากฐานข้อมูลที่ได้มาตรฐานในกิจวัตรประจำวัน ที่มีการควบคุมสภาพแวดล้อมจำนวนบุคคล ที่มีการทับซ้อนกันของวัตถุไม่มากนัก รวมทั้งตำแหน่งกล้องในมุมองศาที่เอื้ออำนวยสำหรับการสกัดคุณลักษณะเริ่มต้นของโครงร่างบนร่างกาย ดังนั้นสำหรับงานวิจัยนี้ คัดเลือกข้อมูลจากฐานข้อมูล UR FALL [22] ข้อมูลถูกสกัดคุณลักษณะตามโครงร่างกระดูกในรูปแบบสองมิติด้วย OpenPose [23] ที่สัมพันธ์กับตำแหน่งข้อต่อสำคัญของการเคลื่อนไหวบนร่างกายจากภาพวิดีโอ 25 จุดในแต่ละเฟรม ที่มี 15 จุดสำคัญจากร่างกาย (เริ่มจากตำแหน่ง 0 ถึง

14) ดังแสดงในรูปที่ 2 ก. จะถูกสกัดออกเป็นข้อมูลตามตำแหน่งร่างกายรวมทั้งมุมองศาที่ได้จากตำแหน่งสำคัญดังแสดงในรูปที่ 2 ข. ข้อมูลถูกจัดเก็บลงในตัวแปร $B = \{B_1, B_2, \dots B_j\}$ กำหนดให้ค่า $B_j \in \mathbb{R}^{\omega \times h}, j \in \{1 \dots J\}$ แทนจำนวนรวมทั้งหมดของร่างกายทุกส่วน แต่ละส่วนของร่างกายเป็นชุดข้อมูลแทนด้วยตัวแปร $P = \{P_1, P_2, \dots, P_C\}$ กำหนดให้ค่าเวกเตอร์ในตำแหน่งต่างๆของร่างกายทั้งซ้าย และขวา สามารถเขียนแทนด้วย $P_c \in \mathbb{R}^{\omega \times h \times 2}, c \in \{1 \dots C\}$ เมื่อ C แทนจำนวนของแขนขาทั้งหมด



ก.กำหนดตำแหน่งบนโครงร่างกระดูก ข.กำหนดมุมบนโครงร่าง
รูปที่ 2 แบบจำลองแสดงโครงร่างกระดูก OpenPose [23]

3.2 การสร้างแนวการเคลื่อนที่บนส่วนร่างกาย

การคำนวณคุณลักษณะท่าทางการเคลื่อนไหวจากตำแหน่งของจุดบนแต่ละเฟรม สามารถสร้างความสัมพันธ์ระหว่างจุดเพื่อหาค่าของมุมบนเส้นที่มีการแบ่งส่วน มีความสัมพันธ์ระหว่างมุมองศาดังแสดงในตารางที่ 1 โดยกำหนดให้เวกเตอร์ $\vec{v}$ ที่มาจากจุดสองจุดกำหนดให้เป็น $p_i = (x_i, y_i)$ และ $p_j = (x_j, y_j)$ เขียนเป็นสมการได้

$$\vec{v} = (x_j - x_i, y_j - y_i) \quad (1)$$

มุม θ ระหว่างสองจุดที่เชื่อมต่อกันบนโครงร่างมนุษย์จะถูกแทนด้วย $\vec{v}$ และ $\vec{u}$ เขียนเป็นสมการได้

$$\theta = \arccos((\vec{v} * \vec{u})/\|\vec{u}\|\|\vec{v}\|) \quad (2)$$

เมื่อ $\|\vec{v}\|$ แทนค่าความยาวของเวกเตอร์ $\vec{v}$ ดังนั้นแต่ละเฟรมประกอบด้วยคุณลักษณะของมุมที่มีทั้งหมด 14 ค่าจากส่วนต่างๆของร่างกาย สามารถเขียนเป็นสมการได้ดังนี้

$$\beta = (\theta_1, \theta_2, \dots, \theta_{14}) \quad (3)$$

กำหนดทิศทางการเคลื่อนที่ ของโครงร่างกระดูก เมื่อถูกตรวจจับบนภาพวิดีโอ จะวัดจากจุดเริ่มต้นจนถึงจุดสุดท้ายที่มีการเคลื่อนที่ไปเขียนเป็นสมการได้ดังนี้

$$T = (\Delta P_t, \dots, \Delta P_{t+F-1}) \quad (4)$$

เมื่อกำหนดให้ค่าเวกเตอร์การกระจัด (displacement vector)

$$\Delta P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t) \quad (5)$$

ดังนั้นทำการนอร์มอลไลซ์ผลรวมของขนาด (sum of the magnitudes) ของเวกเตอร์การกระจัด เขียนเป็นสมการได้

$$T' = T / \sum_{j=t}^{t+F-1} \|\Delta P_j\| \quad (8)$$

ดังนั้น เมื่อมีการคำนวณทุกๆจุดบนเฟรม สามารถเขียนเป็นสมการได้ดังนี้

$$T = (T'_1, T'_2, \dots, T'_{15}) \quad (9)$$

ตารางที่ 1. แสดงความสัมพันธ์ระหว่างส่วนของร่างกายที่ 1 และ 2 เพื่อแสดงตำแหน่งของมุมที่ i

มุมที่ (θ_i)	ส่วนร่างกายที่ 1	ส่วนร่างกายที่ 2
1	ร่างกายหลัก	ไหล่ซ้าย
2	ร่างกายหลัก	ไหล่ขวา
3	ร่างกายหลัก	สะโพกซ้าย
4	ร่างกายหลัก	สะโพกขวา
5	ไหล่ซ้าย	คอ
6	ไหล่ขวา	คอ
7	ต้นแขนซ้าย	แขนซ้าย
8	แขนซ้าย	ไหล่ซ้าย
9	ต้นแขนขวา	แขนขวา
10	แขนขวา	ไหล่ขวา
11	ต้นขาซ้าย	สะโพกซ้าย
12	ต้นขาซ้าย	ขาซ้าย
13	ต้นขาขวา	สะโพกขวา
14	ต้นขาขวา	ขาขวา

4. การเรียนรู้เชิงลึกด้วยเทคนิคโครงข่ายประสาทเทียม

การสร้างแบบจำลองข้อมูลการเรียนรู้ท่าทางมนุษย์จากการเคลื่อนไหว มีการใช้ข้อมูลตั้งต้นจำนวนมาก ประกอบด้วยข้อมูลจากตำแหน่งสำคัญบนร่างกาย 15จุด รวมทั้งมุมที่สัมพันธ์กับตำแหน่ง 14 ค่า และทิศทางการเคลื่อนที่ ของโครงร่างกระดูกจากเฟรมไปยังเฟรมถัดไป เป็นข้อมูลตั้งต้นสำหรับการเรียนรู้ด้วยโครงข่ายประสาทเทียมแบบสังวัตนาการร่วมกับหน่วยความจำระยะสั้นแบบยาว ดังนั้นการนำโครงข่ายปราสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาวที่สามารถเรียนรู้ลำดับของข้อมูลได้ดี มาทำงานร่วมกับโครงข่ายประสาทเทียมแบบสังวัตนาการ โดยใช้ข้อมูลตั้งต้นเข้าโครงข่ายประสาทเทียมแบบสังวัตนาการ เพื่อเป็นการแยกคุณลักษณะ CNN เป็นอินพุตของโครงข่ายปราสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาว ดังแสดงรายละเอียดไว้ดังนี้

4.1 โครงข่ายประสาทเทียมแบบสังวัตนาการ

โครงข่ายประสาทเทียมแบบสังวัตนาการ (Convolutional Neural Networks: CNN) [24] เป็นเพอร์เซ็ปตรอนหลายชั้น (Multi-layer Perceptron) ที่มีโครงสร้างแบบลำดับขั้น (Hierarchical Architectures) ถูกออกแบบมาสำหรับการเรียนรู้ข้อมูลแบบก้าวหน้าในระดับสูงเพื่อใช้ในการจำแนกข้อมูลให้ได้ผลลัพธ์ในขั้นสุดท้าย ข้อดีของการเรียนรู้เชิงลึกจากโครงข่ายประสาทเทียมเพอร์เซ็ปตรอนหลายชั้นคือ สามารถทำการเรียนรู้ข้อมูลแบบอนุกรมเวลาได้ตามประเภทของกิจกรรมที่มีการกำหนดไว้ เพื่อสร้างชุดข้อมูลเริ่มต้นสำหรับการอนุมานในขั้นที่ 1 เขียนเป็นสมการได้ดังนี้

$$B^n = \delta^1(X, S^{n-1}, P^{n-1}), \forall n \geq 2, \quad (10)$$

กำหนดให้เป็น $B^1 = \delta^1(X)$, สำหรับการแทนค่าข้อมูลชุดข้อมูลภาพลงในตัวแปร X เป็นลำดับแรก เพื่อสร้าง Confidence map สำหรับการอนุมาน CNN ในขั้นที่ 1 สามารถทำการเขียนสมการของการทำนายได้ดังนี้

$$P^n = \phi^1(X, S^{n-1}, P^{n-1}), \forall n \geq 2, \quad (11)$$

เมื่อ $P^1 = \phi^1(X)$, และ δ^n และ ϕ^n เป็น CNN สำหรับการอนุมานในขั้นที่ n และการกำหนดค่า Confidence map เพื่อสร้างความสัมพันธ์ระหว่างส่วนต่างๆบนร่างกาย ในการทำนายจะใช้ค่าของ Loss function ของทุกขั้นตอนทำนายร่วมกับค่า ground truth เพื่อกำหนดค่าต่างๆเก็บไว้ ดังนั้นสามารถเขียนค่า Loss function ในแต่ละส่วนบนร่างกายได้ดังนี้

$$\mathcal{L}_B^t = \sum_{j=1}^{J} \sum_p W(p). \left\| S_j^t(p) - S_j^*(p) \right\|_2^2, \quad (12)$$

เมื่อกำหนดให้ S_j^* เป็นค่า Confidence map ของค่า ground truth ในส่วนต่างๆ

$$\mathcal{L}_L^t = \sum_{c=1}^{C} \sum_p W(p). \left\| P_j^t(p) - P_j^*(p) \right\|_2^2, \quad (13)$$

และ P_j^* เป็นค่า ground truth ของเวกเตอร์ของความสัมพันธ์, W แทนค่าเป็นไบนารีมาสก์ (binary mask) เมื่อกำหนดให้ $W(p) = 0$ และ p แทนการให้ค่าที่ผิดพลาดบนตำแหน่งภาพ ดังนั้นทุกภาพจะถูกทำนายด้วยค่า

$$\mathcal{L} = \sum_{t=1}^{T} (\mathcal{L}_B^t + \mathcal{L}_L^t). \quad (14)$$

สำหรับการประเมินค่า $\mathcal{L}_B$ ระหว่างการเรียนรู้ในสมการที่ 12 จะสร้างค่าจาก ground truth ของ confidence map ของ B^*โดยใช้ค่าในตำแหน่งสำคัญที่ได้จากการกำหนด ซึ่งแต่ละ confidence map เป็นการแทนค่าของสองมิติที่ถูกแทนเป็นข้อมูลแต่ละส่วนบนร่างกายในแต่ละพิกเซล จากการทดลองในฐานข้อมูลกำหนดให้มีข้อมูลที่สามารถเขียนเป็นสมการเพื่อหาค่า Confidence map ของ

$$B_j^*(p) = \exp\left(-\frac{\|p - g_j\|_2^2}{\partial^2} \right), \quad (15)$$

กำหนดให้ $g_j \in \mathbb{R}^2$ เป็นตำแหน่งของข้อมูลใน ground truth แต่ละส่วนของร่างกายเป็น j และค่า $p \in \mathbb{R}^2$ ใน B_j^* เมื่อ ∂ เป็นค่าควบคุมค่าสูงสุด

$$B_j^*(p) = \max(0, B_j^*(p)) \quad (16)$$

4.2 โครงข่ายปราสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาว

โครงข่ายประสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory :LSTM) [25] เป็นระบบโครงข่ายประสาทชนิดหนึ่งในประเภทระบบโครงข่ายประสาทเทียมแบบวนซ้ำ (Recurrent Neural Network) ที่มีวิธีการสามารถจดจำข้อมูลได้ในช่วงระยะเวลายาวนานกว่าในรูปแบบอื่น

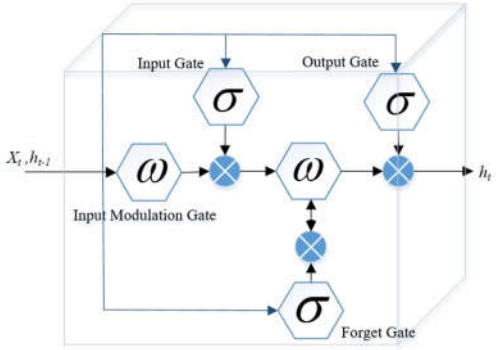
โครงสร้างของ LSTM มีลักษณะเซลที่ประกอบด้วยหลายพารามิเตอร์ในแต่ละ gate จะทำการควบคุมพฤติกรรมของเซลหน่วยความจำ ซึ่งในแต่ละเซลจะถูกควบคุมด้วยฟังก์ชันกระตุ้น ดังแสดงในรูปที่ 3 แสดงรูปแบบของเซลหน่วยความจำที่มีระเวลาในแต่ละขั้น แต่ละ gate มีการทำงานแตกต่างกันดังนี้

1. Input gate (i) เป็นตัวควบคุมข้อมูลเข้าเพื่อทำการพิจารณาชุดข้อมูล (d_i)
2. Forget gate (f) เป็นตัวที่อนุญาตให้ LSTM ลืมความจำก่อนหน้า (c_{t-1}) เพื่อเตรียมพื้นที่สำหรับรับข้อมูลใหม่ด้วยการตัดสินใจจากข้อมูลนำเข้าร่วมกับ hidden state (h_t) ก่อนหน้า
3. Output gate (o) เป็นตัวที่ทำการตัดสินใจว่าข้อมูลในหน่วยความจำจะถูกส่งไปยัง hidden state (h_t)

สามารถเขียนเป็นสมการได้ดังนี้

$$i_t = \sigma(\omega_{di} d_t + \omega_{hi} h_{t-1}) \quad (15)$$

$$f_t = \sigma(\omega_{df} d_t + \omega_{hf} h_{t-1}) \quad (16)$$



รูปที่ 3. โครงข่ายประสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาว

$$o_t = \sigma(\omega_{do} d_t + \omega_{ho} h_{t-1}) \quad (17)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \emptyset(\omega_{dc} d_t + \omega_{hc} h_{t-1}) \quad (18)$$

$$h_t = o_t \odot \phi(c_t) \quad (19)$$

กำหนดให้ σ แทน ฟังก์ชันซิกมอยด์ (sigmoidal function), $\emptyset$ แทน ไฮเบอรโบลิกเทนเจนท์ (hyperbolic tangent), และ c แทน ฟังก์ชันกระตุ้น (activation vector) สำหรับข้อมูลค่าน้ำหนักระหว่างข้อมูลนำเข้าหรือข้อมูลซ่อน (hidden input) ถูกแทนด้วย $\omega_{di}, \omega_{hi}, \omega_{df}, \omega_{hf}, \omega_{do}, \omega_{ho}, \omega_{dc}$ และ ω_{hc} ดังนั้นการปรับปรุงค่าระหว่างขั้นตอนการทำงานจะเป็นการรวมข้อมูลจากขั้นตอนต่างๆเพื่อประกอบการพิจารณาในการปรับปรุงค่าของเซล ซึ่งข้อดีของการใช้ LSTM คือสามารถบอกได้ว่าเมื่อใดจะทำการปรับปรุงเซล อ่านเซลหรือเขียนข้อมูลเข้ามาซึ่งตัว gate จะเป็นตัวควบคุมการไหลของข้อมูล ในรูปที่ 3 แสดงโครงสร้างของ LSTM

5. การวัดประสิทธิภาพและการประเมินผลการทดลอง

งานวิจัยนี้ได้นำเสนอวิธีการการเรียนรู้ท่าทางมนุษย์จากการเคลื่อนไหวด้วยเทคนิค CNN-LSTM โดยใช้ OpenPose [23] เป็นเครื่องมือในการสกัดข้อมูลตำแหน่งข้อต่อจากร่างกายมนุษย์ที่เคลื่อนไหวในสภาพแวดล้อมที่แตกต่างกัน และยังสามารถทำงานได้ดีแม้กระทั่งสภาพแวดล้อมที่มีแสงสว่างน้อย สำหรับการทดลองเพื่อทดสอบวิธีการและทฤษฎีเบื้องต้นได้ทำการทดลองในระบบควบคุมโดยการได้คัดเลือกข้อมูลภาพกิจวัตรประจำวันของมนุษย์ (Activities of Daily Living) [15] จากฐานข้อมูล UR Fall [22] ที่ได้มาตรฐานประกอบด้วยชุดข้อมูล 70 กิจกรรมในชีวิตประจำวันที่มีอัตราความเร็วของวิดีโอ 30 เฟรมต่อวินาที ขนาดภาพ 640x480 พิกเซล ดังนั้นชุดข้อมูลที่นำมาใช้ทดลองจะถูกคัดเลือกตามสภาพแวดล้อมและท่าทางมนุษย์ที่เหมาะสม ประกอบด้วยสภาพแวดล้อม 10 สถานที่ภายในภาพประกอบด้วย กลุ่มตัวอย่างที่มีทั้งชายและหญิงที่มีปฏิสัมพันธ์กันแบบไม่ซับซ้อน สามารถจัดกลุ่มเป็นท่าทางกิจกรรมพื้นฐาน ได้ดังนี้ (1) ดังนี้ ยืน (standing), นั่ง (sitting), หมอบลง (crouching down), เก็บวัตถุ (picking up), ล้มลง (falling down) และ นอน (lying down) ดังแสดงตัวอย่างจากฐานข้อมูลในรูปที่ 4 การทดลองกำหนดให้เป็นชุดเรียนรู้ 80% และชุดทดสอบ 20% ข้อมูล และทำการทดลองโดยสุ่มข้อมูลออกเป็น 2 ชุดข้อมูลสำหรับ 6 กลุ่มกิจกรรม และ โดยแต่ละชุด

ข้อมูลจะมีขนาดความยาว 30 และ 60 เฟรม ประกอบด้วย ข้อมูล Synchronization, ข้อมูล Accelerometer, ข้อมูลทิศทางการเคลื่อนที่ และตำแหน่งข้อต่อสำคัญ 15 จุด กำหนดค่าพื้นฐานสำหรับ CNN ขนาดของตัวกรอง (filter) เพื่อดึงคุณลักษณะไว้ที่ 3 x 3 stride การเลื่อนของตัวกรองเป็น 1 และค่าชั้นพลูลิ่ง (pooling) เป็น 3 และทำการปรับค่าพารามิเตอร์ค่าน้ำหนักสำหรับการเรียนรู้ที่รวดเร็ว เพื่อนำไปเปรียบเทียบและทำการสร้างแบบจำลองและวัดประสิทธิภาพด้วยวิธีการโครงข่ายประสาทเทียมแบบสังวัตนาการร่วมกับโครงข่ายประสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาว (Convolutional neural network and long short-term memory network: CNN-LSTM) ที่นำเสนอเพื่อเปรียบเทียบกับวิธี โครงข่ายประสาทเทียมแบบสังวัตนาการ (Convolutional Neural Networks: CNN) และโครงข่ายประสาทเทียมแบบสังวัตนาการ-ซัพพอร์ตเวกเตอร์แมชชีน (Convolutional Neural Network-Support Vector Machines: CNN-SVM) โดยมีการประเมินประสิทธิภาพด้วยค่าความถูกต้องของแต่ละกลุ่มข้อมูล ประกอบด้วย ค่าความแม่นยำ (Precision: Pre.) และ ค่าความระลึก (Recall.) ดังแสดงผลการทดลองในตารางที่ 2 ถึง ตารางที่ 5 ตามลำดับ

จากตารางที่ 2 และ 3 แสดงผลการจำแนกข้อมูลกิจกรรมด้วยค่าความแม่นยำ (Pre.) และ ค่าความระลึก (Recall) ของชุดข้อมูลที่ 1 โดยกำหนดค่าของโหนดซ่อนเป็น 128 และ 256 ตามลำดับ จากตารางที่ 2 กิจกรรม Falling down มีค่าความแม่นยำสูงที่สุดอยู่ที่ 74.8%, 75.5% และ 76.6% จากวิธี CNN, CNN-SVM และ CNN-LSTM ตามลำดับ แต่เมื่อมีการเพิ่ม

ตารางที่ 2. เปรียบเทียบประสิทธิภาพของการจำแนกท่าทางกิจกรรมชุดข้อมูลที่ 1 ด้วยโหนดซ่อน 128

Categories	Performance (%)					
	CNN		CNN-SVM		CNN-LSTM	
	Pre.	Recall	Pre.	Recall	Pre.	Recall
Standing	71.0	72.4	74.5	76.0	78.7	79.5
Sitting	72.6	68.3	75.3	68.6	76.0	76.8
Crouching down	63.2	71.3	65.2	75.0	70.7	70.7
Picking-up	69.1	67.1	69.7	71.0	72.3	73.1
Falling down	74.8	64.2	75.5	67.0	76.6	72.6
Lying down	68.2	73.5	72.8	75.0	75.8	78.1
Average Accuracy	69.64		71.95		75.04	

ตารางที่ 3. เปรียบเทียบประสิทธิภาพของการจำแนกท่าทางกิจกรรม ชุดข้อมูลที่ 1 ด้วยโหนดซ่อน 256

Categories	Performance (%)					
	CNN		CNN-SVM		CNN-LSTM	
	Pre.	Recall	Pre.	Recall	Pre.	Recall
Standing	79.0	79.8	79.0	72.3	85.2	88.2
Sitting	79.5	72.9	76.9	73.7	77.5	74.5
Crouching down	67.6	74.3	71.4	77.3	77.0	77.0
Picking-up	70.9	70.9	73.8	78.4	75.4	76.9
Falling down	76.7	68.7	78.6	70.6	79.6	76.6
Lying down	70.8	76.2	75.8	73.5	77.7	80.0
Average Accuracy	73.9		75.1		78.8	

จำนวนโหนดซ่อนเป็น 256 ในตารางที่ 3 แสดงกิจกรรม Falling down จะมีค่าความแม่นยำเฉลี่ยเพิ่มขึ้นเฉลี่ย 2.67% แต่ค่าความแม่นยำในกิจกรรม Standing มีค่าสูงสุดอยู่ที่ 85.2% และถัดมาจะเป็นกิจกรรม Lying down 77.7% ที่จำแนกด้วยเทคนิค CNN-LSTM สำหรับกิจกรรม Crouching down และ Picking up มีค่าความแม่นยำเป็น 67.6% และ 70.9% และเพิ่มขึ้นเป็น 77% และ 75.4%สำหรับ การทดลองด้วย CNN และ CNN-LSTM

จากตารางที่ 4 และ 5 แสดงผลการทดลองด้วยชุดข้อมูลที่ 2 แบ่งตามจำนวนโหนดซ่อน 128 และ 256 ตามลำดับ จะสังเกตเห็นว่ากิจกรรม Standing และ Lying down มีค่าความแม่นยำอยู่ที่ 72.5%, 70.3% และค่าความระลึกอยู่ที่ 77.6% และ 73.6% สำหรับการจำแนกด้วย CNN และเมื่อจำแนกด้วย CNN-LSTM มีค่าความแม่นยำอยู่ 80.8% และ 81.2%

ตารางที่ 4. เปรียบเทียบประสิทธิภาพของการจำแนกท่าทางกิจกรรมชุดข้อมูลที่ 2 ด้วยโหนดซ่อน 128

Categories	Performance (%)					
	CNN		CNN-SVM		CNN-LSTM	
	Pre.	Recall	Pre.	Recall	Pre.	Recall
Standing	72.5	77.6	75.4	74.7	80.8	79.2
Sitting	69.8	61.2	72.5	71.8	75.2	75.2
Crouching down	63.5	62.2	74.5	73.1	74.5	76.7
Picking-up	65.0	64.4	69.1	73.2	75.0	72.9
Falling down	67.9	70.5	71.1	71.8	75.2	78.1
Lying down	70.3	73.6	76.1	74.1	81.2	79.6
Average Accuracy	68.2		73.1		77.0	

ตารางที่ 5. เปรียบเทียบประสิทธิภาพของการจำแนกท่าทางกิจกรรมชุดข้อมูลที่ 2 ด้วยโหนดซ่อน 256

Categories	Performance (%)					
	CNN		CNN-SVM		CNN-LSTM	
	Pre.	Recall	Pre.	Recall	Pre.	Recall
Standing	76.2	72.6	79.0	75.2	83.3	82.1
Sitting	75.2	71.2	80.8	76.9	81.7	78.3
Crouching down	70.4	70.7	76.5	80.6	79.5	81.4
Picking-up	75.0	76.6	77.4	78.3	79.2	81.6
Falling down	71.4	77.3	75.5	79.8	79.8	79.8
Lying down	76.3	71.2	77.2	75.3	82.6	81.8
Average Accuracy	73.7		77.7		80.9	

และค่าความระลึกอยู่ที่ 79.2%, 79.6% ตามลำดับ ด้วย 128 โหนดซ่อน แสดงในตารางที่ 4 และเมื่อทำการทดลองด้วยการเพิ่มจำนวนโหนดซ่อนเป็น 256 ในตารางที่ 5 ได้ค่าความแม่นยำ 83.3%, 82.6% และ 81.7% ในกิจกรรม Standing, Lying down และ Sitting ด้วยเทคนิค CNN-LSTM ตามลำดับ จากการทดลองจะเห็นว่ามีค่าความแม่นยำเพิ่มขึ้นเมื่อจำนวนโหนดซ่อนเพิ่มขึ้น และได้แสดงการเปรียบเทียบตามชุดการทดลองในแต่ละเทคนิคในตารางที่ 6 เมื่อทำการทดลองด้วยชุดข้อมูลที่ 1 ด้วยโหนดซ่อน 128 ได้ค่าเฉลี่ยความถูกต้องของ CNN, CNN-SVM และ CNN-LSTM เป็น 69.6%, 72% และ 75% ตามลำดับและเมื่อมีการเพิ่มจำนวนโหนดซ่อนค่าเฉลี่ยความถูกต้องที่ได้

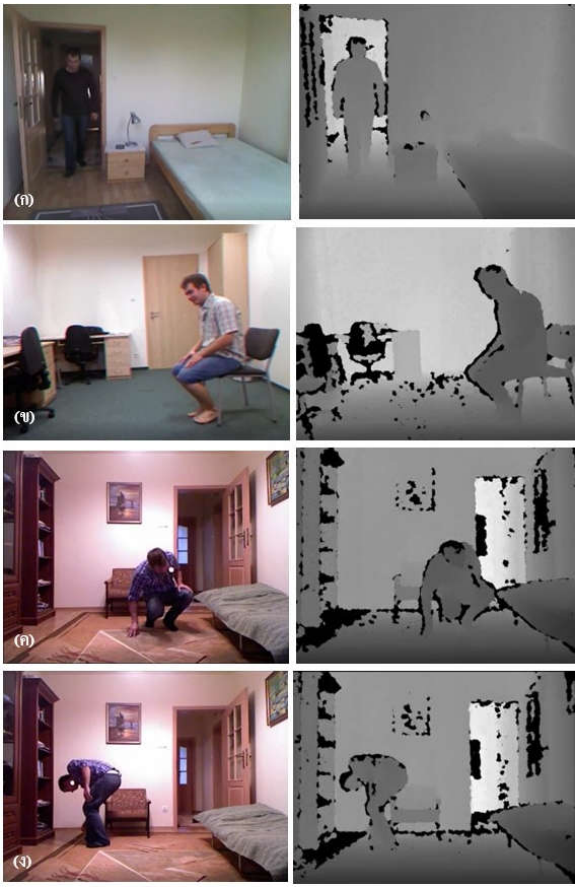
ตารางที่ 6. เปรียบเทียบประสิทธิภาพของการจำแนกท่าทางกิจกรรมระหว่างชุดข้อมูล

Data Set	Nodes	Model Average Accuracy and Number of Nodes (%)		
		CNN	CNN-SVM	CNN-LSTM
I	128	69.6	72.0	75.0
	256	73.6	74.2	78.6
II	128	68.1	73.1	76.9
	256	74.1	77.4	81.0

เพิ่มขึ้นประมาณ 3.3 % สำหรับชุดข้อมูลแรก และค่าความถูกต้องเฉลี่ยเพิ่มขึ้น 4.8% สำหรับชุดข้อมูลที่สอง และเมื่อมีการใช้เทคนิค LSTM เพิ่มขึ้นจะทำให้ค่าความถูกต้องเพิ่มขึ้นโดยเฉลี่ยจากวิธี CNN-SVM ถึง 3.7% และจากวิธี CNN 6.5% สามารถสรุปได้ว่าเมื่อมีการใช้เทคนิค LSTM ร่วมกับ CNN ด้วยการเพิ่มจำนวนโหนดซ่อน สามารถช่วยในการคัดเลือกคุณลักษณะจาก CNN เข้ามายังลำดับชั้น LSTM เพื่อจำแนกท่าทางได้ด้วยค่าเฉลี่ยสูงสุดถึง 81% สำหรับชุดข้อมูลที่ 2 ด้วยโหนดซ่อน 256 โหนด

6. สรุปผลการทดลอง

งานวิจัยนี้ได้นำเสนอการสร้างแบบจำลองท่าทางมนุษย์ด้วยโครงข่ายประสาทเทียมแบบสังวัตนาการร่วมกับโครงข่ายปราสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาวเพื่อจำแนกท่าทางมนุษย์จากการเคลื่อนไหวในกิจกรรมชีวิตประจำวัน เพื่อให้ได้ผลลัพธ์ตามกลุ่มความหมายที่ดีที่สุด ได้มีการใช้ข้อมูลตำแหน่งข้อต่อสำคัญร่วมกับข้อมูลทิศทางการเคลื่อนที่บนร่างกายสามารถ แบ่งกลุ่มตามท่าทางพื้นฐาน 6 กลุ่ม ดังนี้ ยืน, นั่ง, หมอบลง , เก็บวัตถุ, ล้มลง และ นอน จากผลการทดลองในการเรียนรู้ด้วยการกำหนดโหนดซ่อนเป็น 256 จะเห็นว่า มีค่าเฉลี่ยความถูกต้องของการจำแนกท่าทางสูงถึง 81% สำหรับการทดลองในชุดข้อมูลที่ 2 และ78.6% สำหรับการทดลองด้วยชุดข้อมูลที่ 1 แต่อย่างไรก็ตามเมื่อทำการทดลองนั้นยังมีข้อจำกัดของสภาพแวดล้อมที่เปลี่ยนแปลง และจำนวนคนพร้อมวัตถุที่ร่วมเฟรม การทับซ้อนกันของวัตถุเป็นการยากในการสกัดตำแหน่งข้อมูลบนโครงร่างกระดูก ทำให้ยังคงเป็นสิ่งที่ต้องแก้ไขและปรับปรุงวิธีการเพื่อให้ตอบรับกับข้อมูลที่หลากหลายมากขึ้นสำหรับการทดลองต่อไป



(ก) Standing (ข) Sitting (ค) Crouching down (ง) Picking up

รูปที่ 4. ตัวอย่างภาพ RGB และภาพระดับเทา จากฐานข้อมูล UR Fall [22]

เอกสารอ้างอิง

[1] C. Galleguillos and S. Belongie, “Context Based Object Categorization: A Critical Survey”, Computer Vision and Image Understanding, vol. 114, pp. 712-722, 2010.

[2] B. Luo, Y. Sun, G. Li, D. Chen, and Z. Ju, “Decomposition algorithm for depth image of human health posture based on brain health,” Neural Computing and Applications, vol. 32, no. 10, pp. 6327–6342, 2020.

[3] J. A. Graham, and M. Argyle, “A cross-cultural study of the communication of extra-verbal meaning by gestures”, International Journal of Psychology, vol 10, no.1, pp. 57–67, 1975.

[4] A. Wilson, Paul and Barbara Lewandowska-Tomaszczyk, “Affective Robotics: Modelling and Testing Cultural Prototypes,” Cognitive Computation, vol 6, no.4. pp. 814, 2014.

[5] S. Vishwakarma and A. Agrawal, “A survey on activity recognition and behavior understanding in video surveillance,” The Visual Computer, vol. 29, no.10, pp. 983-1009, 2013.

[6] H. Mousavi Hondori and M. Khademi. A review on technical and clinical impact of Microsoft kinect on physical therapy and rehabilitation. Journal of Medical Engineering, 2014.

[7] M. A. Jaffar, M. S. Zia, M. Hussain, A. B. Siddiqui, S. Akram, and U. Jamil, “An ensemble shape gradient features descriptor based nodule detection paradigm: a novel model to augment complex diagnostic decisions assistance,” Multimedia Tools and Applications, 2018.

[8] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, “Chapter Sign Language Recognition Using Convolutional Neural Networks,” European Conference on Computer Vision, pp. 572-578, 2015.

[9] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” International Conference on Learning Representations, 2015.

[10] Y. Wang and M. Hoai. “Improving human action recognition by non-action classification,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2698-2707, 2016.

[11] Li J., X. Liu, W. Zhang, M. Zhang, J. Song and N. Sebe, “Spatio-Temporal Attention Networks for Action Recognition and Detection,” IEEE Transactions on Multimedia, 2020.

[12] Thurau, C. and Hlav´a˘c, V., “n-grams of action primitives for recognizing human behavior,” Computer Analysis of Images and Patterns. Springer Berlin / Heidelberg. of Lecture Notes in Computer Science, vol. 4673, pp. 93–100. 2007.

[13] Alexandros Andre Chaaraoui, Pau Climent-Péreza, and Francisco Flórez-Revuelta, “Silhouette-based human

action recognition using sequences of key poses," Pattern Recognition Letters, vol. 34, Issue 15, pp. 1799-1807, 2013.

[14] S. Yi and V. Pavlovic, "Sparse granger causality graphs for human action classification," Proceedings of the 21st International Conference on Pattern Recognition, pp. 3374-3377, 2012.

[15] O. Ojetola, E. Gaura, and J. Brusey, "Data Set for Fall Events and Daily Activities from Inertial Sensors," In Proceedings of the 6th ACM Multimedia Systems Conference, pp. 18–20, 2015.

[16] F. Hussain, M. Ehatisham-ul-Haq, M.A. Azam, "Activity-Aware Fall Detection and Recognition Based on Wearable Sensors," IEEE Sensors Journal, pp. 4528–4536, 2019.

[17] T. Nguyen, M. Cho, and T. Lee, "Automatic Fall Detection Using Wearable Biomedical Signal Measurement Terminal," In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 5203–5206, 2009.

[18] M. S. Alzahrani, S. K. Jarraya, H. Ben-Abdallah, and M. S. Ali, "Comprehensive evaluation of skeleton features-based fall detection from Microsoft Kinect v2," Signal, Image and Video Processing, vol. 13, no. 7, pp. 1431–1439, 2019.

[19] S. Ghazal, U. S. Khan, M. Mubasher Saleem, N. Rashid, and J. Iqbal, "Human activity recognition using 2D skeleton data and supervised machine learning," IET Image Processing, vol. 13, no. 13, pp. 2572–2578, 2019.

[20] A. Franco, A. Magnani, and D. Maio, "A multimodal approach for human activity recognition based on skeleton and RGB data," Pattern Recognition Letters, vol. 131, pp. 293–299, 2020.

[21] Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299, 2017.

[22] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless Accelerometer," Computer Methods Program Biomed, vol 117, pp. 489–501, 2014.

[23] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields", Computer Vision and Pattern Recognition, 2018.

[24] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition," Computer Vision and Pattern Recognition, 2014.

[25] Je Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," Computer Vision and Pattern Recognition, 2014.