Kasem Bundit Engineering Journal Vol.9 No.1 January-April 2019

การจำแนกภาพความสัมพันธ์ของแอ็คชันด้วยวิธีการโครงข่ายประสาทเทียมแบบสังวัตนาการ-ซัพพอร์ตเวกเตอร์แมชชีน

IMAGE CLASSIFICATION WITH RELATIONSHIP OF ACTIONS BY USING CONVOLUTIONAL NEURAL NETWORK-SUPPORT VECTOR MACHINES

นัศพ์ชาณัณ ชินปัญช์ธนะ
อาจารย์, วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์
110/1-4 ถ.ประชาชื่น เขตหลักสี่ กรุงเทพฯ 10210, nutchanun.cha@dpu.ac.th

Nutchanun Chinpanthana
Lecturer, College of Innovative Technology and Engineering,
Dhurakij Pundit University, 110/1-4 Prachachuen rd. Laksi, Bangkok 10210, Thailand,
nutchanun.cha@dpu.ac.th

บทคัดย่อ

กระบวนการจำแนกข้อมูลภาพเป็นหัวข้อที่น่าสนใจในสาขาด้านการประมวลผลภาพ มีนักวิจัยที่พยายามใช้เทคนิคต่างๆ เพื่อศึกษาและแก้ไขปัญหาที่ ดังนั้นในงานวิจัยนี้ได้สำรวจทฤษฎีและงานวิจัยต่างๆ ในปัจจุบัน ที่เกี่ยวข้องกับระบบการจำแนกภาพ การจำแนกภาพด้วยคำศัพท์ และกระบวนการจำแนกภาพแบบอัตโนมัติ เห็นว่ากระบวนการที่ใช้คำศัพท์ในการจำแนกนั้นอาจจะไม่เพียงพอสำหรับการแทนค่าสำหรับการจำแนกภาพทั้งหมด ดังนั้นในงานวิจัยนำเสนอความสัมพันธ์ของแอ็คชันที่เกิดขึ้นระหว่างวัตถุและแอ็คชันโดยจะแทนด้วย 3 ความสัมพันธ์ประกอบด้วย implied-by, type-of และ mutually exclusive ข้อมูลภาพ มีการแบ่งลำดับขั้นตอนวิธีการเป็น 4 ส่วนหลักดังนี้ (1) การแทนคำศัพท์ลงบนภาพ (2) การกำหนดแอ็คชันบนภาพ (3) การทำนายความสัมพันธ์ (4) การวัดและการประเมินผลการทำงาน โดยทดสอบด้วยข้อมูลจากฐานข้อมูลภาพแอ็คชันและมีการแสดงประสิทธิภาพการทำงานด้วยค่าเฉลี่ยของความแม่นยำถึง 74.4% สำหรับข้อมูลชุด Set II

คำสำคัญ: การจำแนกข้อมูลภาพ, ความสัมพันธ์แอ็คชัน, การประมวลผลภาพ, การจำแนกความสัมพันธ์แอ็คชัน

ABSTRACT

Image retrieval is an active problem in the digital image processing field. A large number of new techniques and systems have researcher involved and attempted to improve the problems. Therefore, we survey the theoretical and empirical contributions in the current decade related to content base image retrieval, keyword annotation, and automatic image retrieval process. The retrieval process of such keyword based approaches is done by keyword searching model. The model is rather rudimentary and it does not specific enough for representing the actual image retrieval. This paper presents a new approach to represent the interactions between object and action. The interaction relationships are including implied-by, type-of and mutually exclusive. The approach is composed of four main phases: (1) Keyword Annotation (2) Define Relationships (3) Relationship Predictions (4) Measurement and Evaluation. We train and test our model on a large scale image dataset of relationship actions. The experimental results indicate that our proposed approach offers significant performance improvements in the classification of relationship actions with maximum success rate of 74.4% in Data Set II.

KEYWORDS: image classification, action relationships, image processing, classification of action relationships

1. บทนำ

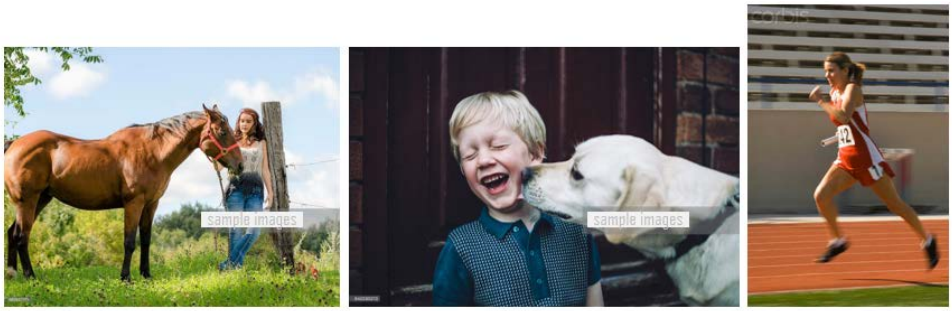
ปัจจุบันความก้าวหน้าของอุปกรณ์อิเลคทรอนิกส์รวมทั้งเทคโนโลยีได้พัฒนาไปอย่างรวดเร็ว และราคาอุปกรณ์ถูกลง ขนาดหน่วยความจำเพิ่มมากขึ้น การจัดเก็บข้อมูลมีขนาดใหญ่ขึ้นตามจำนวนข้อมูลที่เพิ่มขึ้น เช่นเดียวกันกับการพัฒนาการของภาพถ่ายที่มีจำนวนเพิ่มมากขึ้นอย่างรวดเร็วเช่นเดียวกัน ทำให้การจำแนกข้อมูลภาพบนฐานข้อมูลขนาดใหญ่ยังคงเป็นหัวข้องานวิจัยทางด้านการประมวลผลภาพ (Image Processing) ที่มีความท้าทายอย่างมาก ทุกอย่างบนฐานข้อมูลส่วนใหญ่จะมีการเก็บในรูปแบบทั้งข้อความและข้อมูลภาพ และสิ่งที่ยากคือการจำแนกข้อมูลภาพอย่างไรเพื่อให้ได้ข้อมูลภาพตรงตามความหมายของผู้ใช้งานการอย่างแท้จริง มีนักวิจัยพยายามใช้วิธีการเรียนรู้จากข้อมูลตัวอย่าง ทั้งในรูปแบบของการนำข้อมูลระดับต่ำ ที่ประกอบด้วย สี, รูปทรง, ลวดลาย และข้อมูลอื่นๆ เข้ามาใช้ในการค้นคืนภาพที่เรียกว่า Content-Based Image Retrieval (CBIR) [1-3] และได้มีการปรับปรุงพัฒนาในรูปแบบของการใช้คำศัพท์และความสัมพันธ์ภายในภาพเพื่อเป็นข้อมูลสำหรับการจำแนกข้อมูลภาพแทน และถูกเรียกว่า Annotation-Based Image Retrieval (ABIR) [4] แต่อย่างไรก็ตามยังคงมีการพัฒนาโดยนำวิธีการแทนข้อมูลด้วยกราฟ

แบบลำดับขึ้น [5, 6] หรือการวัดความคล้ายด้วยความหมายของคำศัพท์ รวมทั้งวิธีการอื่นที่ซับซ้อนเพื่อนำมาประยุกต์ใช้แต่ยังไม่สามารถได้ภาพผลลัพธ์ตามความต้องการ

2. งานวิจัยที่เกี่ยวข้อง

งานวิจัยสำหรับการจำแนกความหมายภาพเพื่อให้ได้ผลลัพธ์ตามความต้องการของผู้ใช้งานสามารถแบ่งรูปแบบวิธีการค้นคืนความหมายภาพได้ดังนี้ (1) การค้นคืนข้อมูลภาพด้วยข้อมูลระดับต่ำ (CBIR) [1-3] การแทนค่าข้อมูลวัตถุภายในภาพด้วยข้อมูลระดับต่ำ (Low-Level Features) [5, 7] ที่ประกอบด้วย สี, รูปทรง, ลวดลาย หรือข้อมูลความสัมพันธ์อื่นเช่น Scale invariant feature transform หรือ Histograms of oriented gradients [8] ที่นำมารวมกันเพื่อใช้เป็นคุณสมบัติสำหรับการจำแนกข้อมูลภาพ ปัญหาของระบบ CBIR ไม่ว่าจะเป็นส่วนของการทับซ้อนของวัตถุบนภาพหรือความมืดสว่างของรูปภาพที่ถ่าย และบางครั้งวัตถุที่ถ่ายมีการผิดรูปร่างไม่สามารถแปลความหมายได้ ทำให้สิ่งเหล่านี้ยังคงถูกแก้ไขอย่างต่อเนื่องบนระบบ CBIR การนำส่วนของคอมพิวเตอร์วิชัน (Computer vision) การรู้จำแบบ การรู้จำแบบ (Pattern recognition) รวมทั้งการใช้การเรียนรู้ของเครื่อง (Machine learning) เพื่อแก้ไขปัญหาข้างต้น กลุ่มวิจัย Li Fei Fei [9, 10] ได้พัฒนาระบบ CBIR ImageNet ที่ได้รับการยอมรับค่อนข้างมากโดยใช้การผสมผสานการทำงานในส่วนของข้อมูลภาพและการพัฒนาการเรียนรู้ของเครื่องเพื่อทำการค้นคืนภาพให้ได้ตรงตามที่ผู้ใช้งานต้องการ แต่ผลลัพธ์ที่ได้จากการค้นคืนด้วยข้อมูลภาพระดับต่ำจะมีความเหมือนกันทางกายภาพเท่านั้น จะไม่ได้ความหมายภาพที่ต้องการและได้มีการปรับปรุงพัฒนาในรูปแบบของการใช้คำศัพท์เพื่อเป็นข้อมูลสำหรับการค้นคืนแทนถูกเรียกว่า (2) การค้นคืนข้อมูลภาพด้วยข้อมูลคำศัพท์ (ABIR) [11] การแทนค่าข้อมูลวัตถุภายในภาพด้วยคำศัพท์ (keyword) เช่น “sky”, “bicycle”, “home” เป็นต้น โดยทำการให้ความหมายภาพเป็นวัตถุที่เด่นชัดบนภาพดังแสดงในรูปที่ 1 เพื่อใช้สำหรับการค้นคืนข้อมูลภาพแทนข้อมูลระดับต่ำ ผลลัพธ์ที่ได้จะมีความเหมือนกันกับข้อมูลคำศัพท์ที่ใช้ในการค้นหา แต่อย่างไรก็ตาม งานวิจัยส่วนใหญ่พยายามปรับปรุงรูปแบบของผลลัพธ์ของการจำแนกที่ให้มีความหมายตรงตามความต้องการของผู้ใช้งานที่เรียกว่า ความหมายภาพ (Semantic Image) ด้วยการสร้างความสัมพันธ์ของวัตถุภายในภาพ (Spatial Relationships) เช่น “below”, “above”, “hold” เป็นต้น [12] เพื่อหากิจกรรมที่เกิดขึ้นภายในภาพ โดยอาศัยการเรียนรู้เพื่อให้เกิดความสัมพันธ์ของ “person” และ “object” ภายใน แต่อย่างไรก็ตามบางกลุ่มได้พยายามสร้างวิธีการด้วยโครงสร้างความหมายแบบลำดับในเชิงการมองเห็น (Semantic Hierarchy for Vision) [13, 14] กับโครงสร้างของการเรียนรู้รูปแบบเพื่อจำแนกกลุ่มและแบ่งส่วนภาพ เพื่อสร้างความสัมพันธ์ในรูปแบบที่กำหนด และหาความสัมพันธ์ของภาพจากคำศัพท์ภายในด้วยการเรียนรู้ทำให้ครอบคลุมรูปแบบการจำแนกภาพ Deng et al. [15], ใช้กราฟเพื่อสร้างความสัมพันธ์ (DAG relationships) และ Mutual exclusion ระหว่างข้อมูลสำหรับการจำแนกที่ดีขึ้น บางกลุ่ม

พยายามที่จะมีการใช้กลุ่มข้อมูลที่แตกต่างกัน เพื่อทำการเรียนรู้รูปแบบความสัมพันธ์ที่เป็นแบบลำดับขั้นบนพื้นฐานของการวัดความคล้ายและการเกิดขึ้นของความสัมพันธ์ [16] แต่เมื่อมีการเรียนรู้เกิดขึ้นทำให้มีการสร้างองค์ความรู้เพื่อเก็บเป็นข้อมูลคำศัพท์บางที่มาจากเว็บไซต์ที่สร้างขึ้นเพื่อรวบรวมคำศัพท์ที่เกี่ยวข้อง แต่อย่างไรก็ตามรูปแบบของการเรียนรู้แบบอัติโนมัติยังคงมีความจำเป็น



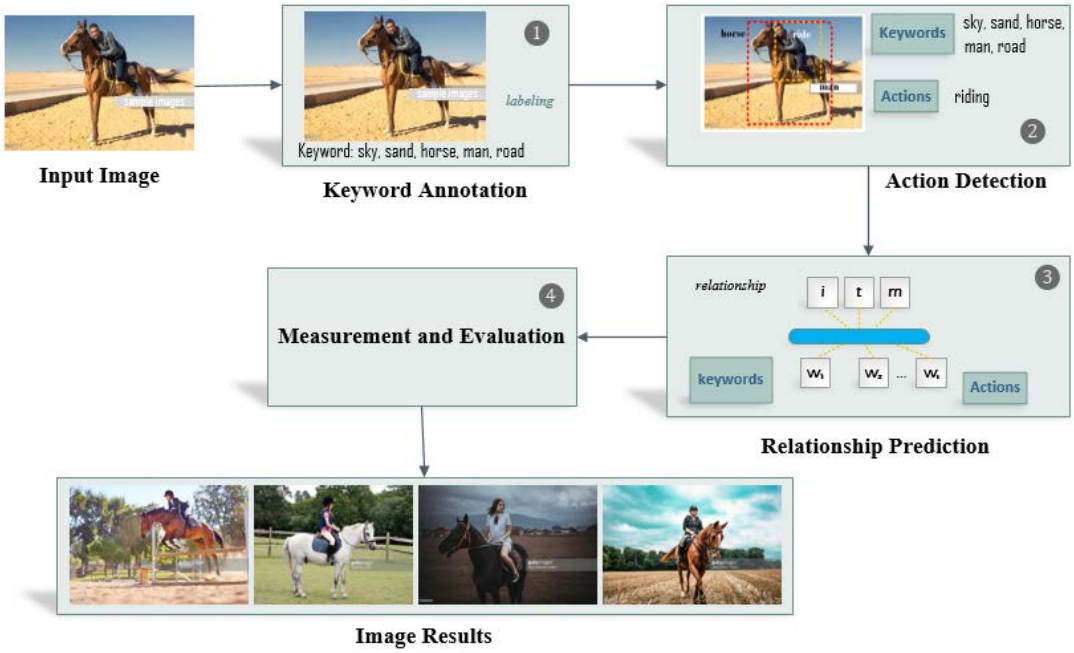
Ground truth	sky, grass, horse, woman	kid, dog	woman
Top 3 key	sky, horse, woman	kid, dog, building	woman, stadium, track
Relationship	(horse, hold, woman)	(kid, kiss, dog)	(woman, run ,track)

รูปที่ 1 ตัวอย่างคำศัพท์ที่ถูกแท็กในฐานข้อมูลภาพ

ปัจจุบันงานวิจัยพยายามที่จะสร้างความสัมพันธ์ภายในภาพ [16, 17] เช่น (girl, on, horse) หรือ (man, eat, apple) เพื่อตอบสนองการแปลความหมายภาพที่มีความซับซ้อนขึ้น ทีมวิจัยของ Cewu Lu และคณะ [13] ได้นำเสนอวิธีการที่ทำการจับคู่วัตถุ หัวเรื่องและความสัมพันธ์ระหว่างกัน ตัวอย่างเช่น รูปภาพที่ประกอบด้วย บุคคล (person), มอเตอร์ไซค์ (motorcycle) และ หมวกกันน็อค (helmet) สามารถสร้างความสัมพันธ์ได้เป็น (person - on – motorcycle), (person -wear – helmet) หรือ (motorcycle - has – wheel) ข้อมูลภาพทั้งหมดถูกรวบรวมและจำแนกด้วยวิธีการ Deep Relational Network จากผลการทดลองจะเห็นว่า การใช้รูปแบบความสัมพันธ์กับวัตถุบนภาพมีความหลากหลาย ทำให้การทำนายความสัมพันธ์ที่เกิดขึ้น จากการทดลองทั้งหมดต้องมีตัวอย่างภาพครอบคลุมมากเพียงพอถึงจะสามารถสื่อถึงความหมายของภาพโดยรวมอย่างแท้จริง ดังนั้นในงานวิจัยนี้จึงได้นำเสนอวิธีการทำนายจำแนกข้อมูลด้วยวิธีการแบบใหม่ด้วยการสร้างพารามิเตอร์ของความสัมพันธ์ระหว่างวัตถุและแอ็คชันจากข้อมูลภายในภาพเพื่อสามารถนำใช้ในการหาภาพผลลัพธ์ที่ตรงกับความต้องการของผู้ใช้งานได้มากขึ้น

3. ขั้นตอนการจำแนกข้อมูลความหมายภาพ

สำหรับในงานวิจัยนี้ได้พัฒนาวิธีการเพื่อให้สามารถจำแนกความหมายภาพให้ได้ตรงกับสิ่งที่ผู้ใช้งานต้องการมากที่สุด จึงได้นำเสนอรูปแบบของการเชื่อมโยงความสัมพันธ์ระหว่างวัตถุและแอ็คชันจากข้อมูลภาพ แบ่งลำดับขั้นตอนวิธีการเป็น 4 ส่วนหลักดังนี้ (1) การแทนคำศัพท์ลงบนภาพ (Keyword Annotation) (2) การตรวจจับแอ็คชันบนภาพ (Action Detection) (3) การทำนายความสัมพันธ์ (Relationship Predictions) (4) การวัดและการประเมินผลการทำงาน (Measurement and Evaluation) ดังแสดงในรูปที่ 2

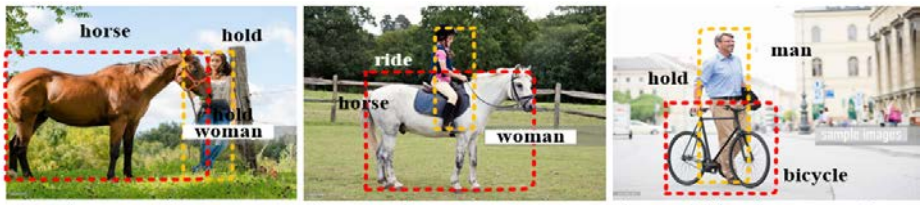


รูปที่ 2 ขั้นตอนการจำแนกความหมายภาพด้วยการทำนายความสัมพันธ์คำศัพท์บนภาพ

3.1 การแทนคำศัพท์ลงบนภาพ

ในขั้นตอนทั้งหมดสิ่งสำคัญที่ช่วยทำให้กระบวนการสร้างการเชื่อมโยงความสัมพันธ์ของความหมายภาพให้มีความสมบูรณ์ขึ้นเพื่อเข้าสู่กระบวนการเรียนรู้อย่างสมบูรณ์เต็มรูปแบบ คือ ขั้นตอนการเตรียมข้อมูล (Data Preprocessing) เริ่มจากคัดเลือกข้อมูลภาพดิจิทัลที่มีวัตถุบนภาพเด่นชัด มีวัตถุภาพพื้นหลังชัดเจน ส่วนภาพมีลักษณะผิดปกติ (Outlier) หรือ คุณลักษณะวัตถุ (Object Characteristic) ไม่ชัดคลุมเครือ มีขนาดวัตถุขนาดเล็กเกินไปไม่สามารถบ่งชี้ชื่อวัตถุได้ หรือภาพถ่ายระยะใกล้ (Close up) จะถูกคัดออก ดังนั้นภาพที่คัดเลือกเข้ามาจะต้องมีความสมบูรณ์ของวัตถุและคำศัพท์ชัดเจน และแปลความหมายภาพนั้นได้อย่างสมบูรณ์

ข้อมูลแอ็คชันบนภาพเกิดจากความสัมพันธ์ของการแอ็คชันของคำนามที่เกิดขึ้น รวมทั้งการลดและจำกัดรูปแบบความสัมพันธ์ (Relationship) เพื่อให้เกิดความรัดกุมของข้อมูลเป็นอีกสิ่งที่จำเป็นเพื่อทำให้ภาพเกิดความหมายเดียวกันมากขึ้น สำหรับแต่ละคำศัพท์การแอ็คชันกำหนดให้มีรูปแบบที่ประกอบด้วยโครงสร้าง คำศัพท์ที่ทำหน้าที่เป็นประธาน หรือคำศัพท์ที่เป็นคำนาม และคำศัพท์ที่เป็นแอ็คชัน (กริยา) ตัวอย่างดังแสดงในรูปที่ 3 “woman holding horse”, “woman riding on horse” หรือ “man holding bike” เป็นต้น วัตถุที่ปรากฏบนภาพ woman horse man และbike มีแอ็คชัน holding และ riding เป็นต้น



Keywords	sky, grass, horse, woman	sky, grass, horse, woman, tree	sky, road, bicycle, man, road
Action	touch	ride	pull

รูปที่ 3 ตัวอย่างคำศัพท์ที่ประกอบด้วยคำนามและแอ็คชันบนฐานข้อมูลภาพ

มีกลุ่มงานวิจัยหลายกลุ่มที่พยายามปรับปรุงวิธีการเพื่อทำการตรวจจับวัตถุ (Object recognition) [12, 17-19] และตรวจจับแอ็คชันบนภาพ (Action recognition) [9, 10, 20-22] ผลการทดลองที่ได้จากการตรวจจับวัตถุและแอ็คชันมีค่าความถูกต้องค่อนข้างดีและสามารถนำวิธีการมาประยุกต์ใช้ได้ ดังนั้นในงานวิจัยนี้ได้เจาะจงนำเสนอวิธีการใหม่เพื่อหาความสัมพันธ์ของวัตถุและแอ็คชั่น จึงได้เลือกวิธีการที่ตรวจจับวัตถุและแอ็คชันบนภาพ พร้อมเลือกฐานข้อมูลที่มีขนาดไม่ใหญ่มาก เพื่อทดสอบการทำงานที่ครอบคลุม จึงเลือกฐานข้อมูลแอ็คชันจาก PASCAL [12] เพื่อเป็นพื้นฐานของการทดลองตามรูปแบบของแอ็คชัน [20-22] ที่กำหนดไว้

3.2 การกำหนดแอ็คชันบนภาพ

ข้อมูลคำศัพท์ภาพ จะถูกแบ่งชนิดเป็น 2 ชนิดดังนี้ คำศัพท์ที่เป็นคำนาม (Noun) และคำศัพท์ที่เป็นแอ็คชัน กำหนดให้ คำศัพท์แอ็คชัน (Action: A) เขียนความสัมพันธ์ของแอ็คชันบนภาพได้ I_A และความสัมพันธ์ระหว่างคำศัพท์การแอ็คชันกำหนดให้เป็น $R_A \subset A$ สำหรับทุกการแอ็คชัน ซึ่งมีค่าน้ำหนักของแอ็คชัน (weight of A : ϖ_A) เป็น $\varpi_A \in R$ และแต่ละภาพ I จะมีข้อมูลคุณลักษณะภาพที่แทนด้วย $f_I \in R$ โดยที่ข้อมูล $f_I = W_{im}CNN(I) + b_{im}$, จัดเก็บข้อมูลในรูปแบบของลำดับขั้นโครงสร้างของโครงข่ายประสาทเทียมแบบสังวัตนาการ (Convolution Neural Network: CNN) [22, 23] เมื่อกำหนดให้ $CNN(I)$ แทนคุณลักษณะของภาพ I และ กำหนดให้พารามิเตอร์

W_{im} และ b_{im}เป็นข้อมูลที่ได้จากการเรียนรู้ระหว่างโมเดลการทำนายความสัมพันธ์ ดังนั้นสามารถกำหนดให้ ค่าน้ำหนักของแอ็คชันสูงสุดสำหรับภาพที่อยู่ในกลุ่มความหมายเดียวกันจะมีค่าเป็นค่าบวก (Positive image) และจะมีการเปรียบเทียบกับค่าที่เป็นลบ ดังนั้นทำให้เกิดการจัดลำดับของการเปรียบเทียบขึ้นและสามารถเขียนเป็นสมการของการจัดลำดับ (C_{ac})ได้ดังสมการที่ (1)

$$C_{ac} = \sum_A \sum_{\substack{I^+ \in I_A \\ I^- \in I_A}} \max(0, 1 + \varpi_A^T (f_{I^-} - f_{I^+})) \quad (1)$$

เมื่อข้อมูลคุณลักษณะภาพ (f_I) แทนภาพความหมายเดียวกัน (I^+) และภาพที่มีความหมายต่างกัน (I^-) แทนด้วย $f_I = (f_{I^-}, f_{I^+}\} \in R$

4. การทำนายความสัมพันธ์

การทำนายความสัมพันธ์ของภาพเป็นการนำความสัมพันธ์ระหว่างวัตถุและแอ็คชันที่เกิดขึ้นบนภาพ มาสร้างความสัมพันธ์ภายใน โดยที่กำหนดคู่ความสัมพันธ์ของคำศัพท์การแอ็คชัน $A, B \subset R_A$ และกำหนดรูปแบบของความสัมพันธ์ที่เกิดระหว่างคำศัพท์ในภาพดังนี้

(1) แบบ implied-by คือ A implied $-$ by B ถ้าเกิดแอ็คชันของ B แสดงให้เห็นว่า สามารถเกิดการแอ็คชัน A ได้ จะมีลักษณะของความสัมพันธ์แบบเดียวกับต้นไม้ กิ่งพ่อแม่ไปยังกิ่งลูก (parent-child node)

(2) แบบ type-of คือ A type $-$ of B แอ็คชันของ A เป็นชนิดหนึ่งของการเกิดการแอ็คชัน B ได้จะมีลักษณะของความสัมพันธ์แบบเดียวกับต้นไม้ กิ่งลูกไปยังกิ่งพ่อแม่ (child-parent node)

(3) แบบ mutually exclusive เมื่อมีการแอ็คชันของ A เกิดขึ้นแล้วจะไม่เกิดขึ้นอย่าง B อย่างแน่นอน แสดงความสัมพันธ์ในรูปที่ 4 เป็นการทำนายความสัมพันธ์จากการเรียนรู้รูปแบบของแอ็คชันที่เกิดขึ้น (ภาพตัวอย่างมีแอ็คชัน: Action A และ Action B) จะเป็นค่าน้ำหนักที่ได้จากความสัมพันธ์ (กลายเป็นค่าน้ำหนักที่เกิดจากแอ็คชัน A: ϖ_A และแอ็คชัน B: ϖ_B) และจะเรียนรู้ตามรูปแบบความสัมพันธ์ทั้ง 3 แบบ สามารถเขียนเป็นสมการความสัมพันธ์ได้ดังนี้ $r_{AB} = [r_{AB}^i, r_{AB}^t, r_{AB}^m] \in [0,3]^3$, เมื่อ r^i, r^t, r^m เป็นความสัมพันธ์ของ Implied-by, Type-of และ Mutually exclusive ตามลำดับ โดยที่ความสัมพันธ์ที่เกิดขึ้นจะใช้โครงข่ายประสามเทียม (Neural network) ในการจัดการความรู้ ด้วยการจัดลำดับชั้นด้วย Softmax Normalization [20, 21] สามารถเขียนเป็นสมการการคาดคะเนความสัมพันธ์ได้ดังสมการที่ (2)

$$r_{AB} = softmax_\beta(\varpi_A \otimes W_\delta^{[1:3]} \otimes \varpi_B + b_\delta), \quad (2)$$

เมื่อกำหนดให้ $W_\delta^{[1:3]} \in \mathbb{R}^{n\times n\times 3}$, n แทนจำนวนข้อมูลของภาพและ $b_\delta \in \mathbb{R}^3$ $softmax_\beta$: $\mathbb{R}^3 \to \mathbb{R}^3$ มีตัวแปร Softmax Normalization เป็น β สำหรับการสร้างความสัมพันธ์ของการแอ็คชัน (Create Action Relationships) เริ่มต้นจากการกำหนดให้สร้างความสัมพันธ์เพื่อใช้ในการเรียนรู้สำหรับการทำนายด้วยคำศัพท์ที่ใช้ในเซตความสัมพันธ์ กำหนดให้ความสัมพันธ์ระหว่างการแอ็คชันของวัตถุมีทั้งหมด 3 ประเภทดังนี้

(1) แบบ A implied $-$ by B เมื่อเวกเตอร์ค่าน้ำหนักการแอ็คชัน (Action Weight Vector: W_A) เป็นอยู่ในลำดับภาพที่อยู่ในกลุ่มภาพ B ที่มีค่ามากกว่าภาพ A ที่ไม่อยู่ในกลุ่มแล้ว กำหนดให้มีค่าความคาดหวังความสัมพันธ์เป็น

$$\mathrm{C}_{\mathrm{AB}}^{\mathrm{i}} = \sum_{I^b\in I_B, I^-\in I_{\bar{A}}} \max(0, 1 + w_A^T(f_{I^-} - f_{I^b})) \quad (3)$$

(2) แบบ A type $-$ of B เมื่อเวกเตอร์ค่าน้ำหนักการแอ็คชันชัน (Action Weight Vector: W_B) อยู่ในลำดับภาพที่อยู่ในกลุ่มภาพ A ที่มีค่ามากกว่าภาพ B ที่ไม่อยู่ในกลุ่มแล้ว กำหนดให้มีค่าความคาดหวังความสัมพันธ์เป็น

$$\mathrm{C}_{\mathrm{AB}}^{\mathrm{t}} = \sum_{I^b\in I_A, I^-\in I_{\bar{B}}} \max(0, 1 + w_B^T(f_{I^-} - f_{I^a})) \quad (4)$$

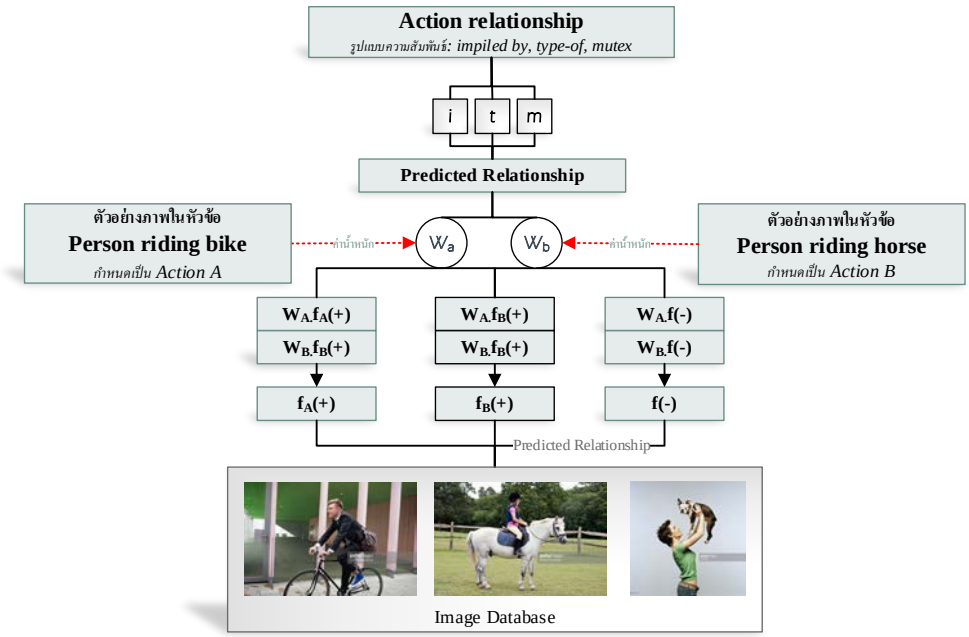
(3) แบบ A mutually exclusive B เมื่อเวกเตอร์ค่าน้ำหนักการแอ็คชันเป็น W_A อยู่ในลำดับภาพที่อยู่ในกลุ่มภาพ A ที่มีค่ามากกว่าภาพ B ที่อยู่ในกลุ่มแล้วจะกำหนดให้มีค่าความคาดหวังความสัมพันธ์เป็น

$$\mathrm{C}_{\mathrm{AB}}^{m} = \sum_{I^a\in I_A, I^-\in I_B} \max(0, 1 + w_A^T(f_{I^b} - f_{I^a})) \quad (5)$$

ดังนั้นทำการรวมสมการที่ (3) ถึง (5) เข้าด้วยการเพื่อสร้างความสัมพันธ์ของการคาดทำนายดังแสดงในสมการที่ 6

$$C_\alpha = \sum_{A\in A, B\in R_A} r_{AB}^i . C_{AB}^i + r_{AB}^t . C_{AB}^t + r_{AB}^m . C_{AB}^m \quad (6)$$

เวกเตอร์ค่าน้ำหนักการแอ็คชัน W_A และ W_B ได้มาจากการเรียนรู้ของฟังก์ชันความสัมพันธ์ที่ดีที่สุดของข้อมูลภาพ ทำให้วิธีการสามารถเลือกความสัมพันธ์ได้อย่างอัตโนมัติเพื่อทำการคำนวณค่าน้ำหนักของการแอ็คชันได้ ดังแสดงตัวอย่างความสัมพันธ์ในรูปที่ 5



รูปที่ 4 ขั้นตอนการสร้างความสัมพันธ์ของแอ็คชัน

รูปที่ 5 ตัวอย่างเซตความสัมพันธ์ของการแอ็คชันของข้อมูลภาพ

5. การวัดและประเมินผลการทำงาน

การวัดและประเมินผลการทำงาน (Measurement and Evaluation) เป็นขั้นตอนสุดท้ายที่สำคัญ เพื่อทดสอบวิธีการข้างต้นว่ามีประสิทธิภาพมากหรือน้อย สอดคล้องกับความต้องการ

เพียงใดเมื่อนำมาใช้งานจริง สำหรับข้อมูลภาพที่ใช้ในการทดลองทั้งหมดจะคัดเลือกภาพที่เห็นวัตถุอย่างเด่นชัดและมีความสอดคล้องกับฐานข้อมูลพื้นฐานจากฐานข้อมูลภาพแอ็คชัน Pascal [12] โดยที่แต่ละกลุ่มคำจะมีแอ็คชันที่มีความสอดคล้องกันทั้งภาพที่อยู่ในกลุ่ม (Positive Image) และภาพที่ไม่อยู่ในกลุ่มซึ่งไม่มีความเกี่ยวข้องกัน (Mutually Exclusive) ในการประเมินนั้นกระทำได้โดยการวัดประสิทธิภาพของการจัดกลุ่มภาพจะถูกพิจารณาเป็นค่าของความถูกต้องของแต่ละกลุ่มข้อมูลสำหรับการทดลองนี้จะใช้ค่าเฉลี่ยของความแม่นยำ (Mean Average Precision: mAP) [19] สำหรับการเปรียบเทียบจะเลือกวิธีการที่สามารถจำแนกได้ดีที่สุดและมีการผสมผสานการเรียนรู้ฟีเจอร์ใหม่ที่เกิดจากข้อมูลสำหรับการทดลองนี้จึงได้เลือกทั้งหมด 3 วิธีการดังนี้

5.1 โครงข่ายประสาทเทียมแบบสังวัตนาการ

โครงข่ายประสาทเทียมแบบสังวัตนาการ (Convolutional Neural Networks: CNN) [16, 17] เป็นเพอร์เซ็ปตรอนหลายชั้น (Multi-layer Perceptron) ที่มีโครงสร้างแบบลำดับขั้น (Hierarchical Architectures) ถูกออกแบบมาสำหรับการเรียนรู้ข้อมูลแบบก้าวหน้าในระดับสูงเพื่อใช้ในการจำแนกข้อมูลให้ได้ผลลัพธ์ในขั้นสุดท้าย สามารถเขียนการดำเนินงานเป็นลำดับชั้นพื้นฐานของ CNN ดังนี้

(1) Convolution Layer ลำดับชั้นจะทำการกรองข้อมูลแบบ 2 มิติ ระหว่างข้อมูลภาพ I, ตัวกรอง w และ h เป็นข้อมูลภาพที่ถูกสร้างขึ้น ให้ CT แทนความสัมพันธ์ของข้อมูลเข้าและออก เมื่อตัวกรองตอบสนองจากข้อมูลเข้าที่เชื่อมต่อกับข้อมูลออกตัวเดียวกันจะถูกเชื่อมต่อกัน สามารถเขียนเป็นสมการ ดังสมการที่ (7)

$$h_j = \sum_{i,k \in CT_{i,k,j}} (I_i * W_k) \tag{7}$$

(2) Sub-sampling Layer เป็นลำดับชั้นที่ถูกสร้างขึ้นเพื่อดำเนินการ “Max-Pooling” เป็นการเปลี่ยนแปลงและลดทอนให้สามารถทำงานได้อย่างรวดเร็วขึ้น

(3) Fully-connected Layer เป็นลำดับชั้นที่รับข้อมูลเข้าสำหรับทุกลำดับชั้นเพื่อทำการรวมเข้าเป็นข้อมูลออก แบบ 1 มิติสำหรับลำดับชั้นถัดไป

(4) Output Layer เป็นลำดับชั้นข้อมูลออกที่ถูกจำแนกเพียงคลาสเดียวแบบ 1 มิติจากลำดับชั้น Fully-connected Layer

โดยปกติทั่วไปขั้นตอนการเรียนรู้ของ CNN จะใช้วิธีการ Gradient Descent Approaches เพื่อลดความผิดพลาดของทุกลำดับชั้นใน CNN ที่มีการเปลี่ยนแปลงตลอดเวลา แต่บางงานวิจัยจะมีการใช้ขั้นตอน Back-propagation สำหรับการเรียนรู้ด้วย Max-Pooling Convolutional Neural Network และใช้ค่าความคลาดเคลื่อนยกกำลังสอง (Sum of the Squared Error) เพื่อลดความ

ผิดพลาด แต่อย่างไรก็ตามเนื่องจากค่าของข้อมูลโดยทั่วไปเป็นแบบไม่เป็นเชิงเส้น (Non-linear) และมีความผิดพลาดมากวิธีการ Stochastic Gradient Descent จึงเป็นวิธีการที่เหมาะสมในการนำมาใช้สำหรับการทดลองเพื่อหลีกเลี่ยงปัญหาที่เกิดการติดขัดเมื่ออยู่ในจุดที่มีค่าต่ำสุด

5.2 ซัพพอร์ตเวกเตอร์แมชชีน

ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines: SVM) [22] เป็นขั้นตอนวิธีการที่สามารถนำมาช่วยวิเคราะห์ข้อมูลและจำแนกข้อมูล โดยใช้หลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่ง (Hyperplane) ที่เป็นเส้นตรงขึ้นมา เพื่อแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกกลุ่มข้อมูลได้ดีที่สุด กำหนดให้ชุดข้อมูลการเรียนรู้เป็นข้อมูลภาพ I ที่มีข้อมูลเป็น $\{y_i, x_i\}_{i=1}^{m}$, เมื่อ $x_i \in R^n$ เป็นรูปแบบข้อมูลเข้าที่ i และ $y_i \in R^n$ เป็นรูปแบบข้อมูลออกที่ i และมีรูปแบบของการจำแนกด้วยวิธีการ SVM ดังสมการ $y(x) = sign(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b)$ เมื่อ α_i เป็นค่าบวกคงที่และ b เป็นค่าคงที่จำนวนจริง โดยปกติแล้ว SVM ถูกนำมาใช้กับข้อมูลที่เป็นเชิงเส้น แต่ในความเป็นจริงแล้วข้อมูลที่นำมาใช้ในระบบการสอนให้ระบบเรียนรู้ส่วนใหญ่มักเป็นข้อมูลแบบไม่เป็นเชิงเส้น ซึ่งสามารถแก้ปัญหาดังกล่าวด้วยการ Kernel Function มาใช้จำแนกข้อมูลบนระนาบหลายมิติ และ $K(\dots)$ แทนด้วยค่าข้อมูลแบบไม่เป็นเชิงเส้น Kernel function ที่ถูกปรับค่าเป็นข้อมูลเข้าที่มีหลายมิติ สามารถเขียนสมการของการลดทอน Cost function ได้ดังสมการที่ (8)

$$minimize_{w,\xi,b}\{\|w\|^2/2 + C\sum_{i=1}^{m}\xi_i\} \quad (8)$$

ภายใต้ข้อจำกัดของ $y_i(w.x_i - b) \geq 1 - \xi_i$ และ $\xi_i \geq 0$, สำหรับ $i = \{1,2,\dots,m\}$ เมื่อ ξ_i เป็นข้อมูลใช้วัดดีกรีการจำแนกข้อมูลที่ผิดพลาดของ x_i และตัวแปรควบคุม C จะควบคุมระหว่างค่าผิดพลาดและค่าสูงสุดของข้อมูลการเรียนรู้ โดยที่ถ้า $C = \infty$ จะนำพาไปยัง Margin ที่ยากของ SVM โดยทั่วไป SVM จะถูกเรียนรู้ด้วยการคูณของค่าลากรองจ์ (Lagrangian) และการแก้ปัญหาด้วยการเขียนโปรแกรมกำลังสอง (Quadratic Programming) แต่อย่างไรก็ตามยังไม่สามารถใช้ได้กับ CNN ทำให้ต้องมีการเรียนรู้ CNN และ SVM พร้อมกันโดยใช้เพียงขั้นตอนเดียวโดยการปรับค่าที่เพิ่มขึ้นแบบออนไลน์ด้วยวิธีการที่เรียกว่า Stochastic Gradient Descent (SGD) [23]

5.3 โครงข่ายประสาทเทียมแบบสังวัตนาการ-ซัพพอร์ตเวกเตอร์แมชชีน

โครงข่ายประสาทเทียมแบบสังวัตนาการ-ซัพพอร์ตเวกเตอร์แมชชีน (Convolutional Neural Network - Support Vector Machines: CNNSVM) เป็นการนำสองวิธีการมาผสมรวมกันโดยที่ CNN จะถูกเรียนรู้ข้อมูลเข้าที่ไม่สัมพันธ์กับรูปแบบของข้อมูล และ SVM ใช้ Kernel Function สามารถจำแนกผลลัพธ์ได้ตามกลุ่มที่ถูกเรียนรู้ได้ ซึ่ง CNN จะสามารถทำงานได้อย่างมีประสิทธิภาพมากสำหรับข้อมูลภาพคงที่แต่ไม่สามารถจำแนกผลลัพธ์ได้เสมอไป เช่นเดียวกันกับวิธี SVM จะถูกจำกัดด้วยค่า Kernel function ที่ซับซ้อนแต่ว่าผลลัพธ์ของการจำแนกข้อมูลที่ดีด้วยวิธีการ Soft-margin Approaches ดังนั้นสามารถเขียนวิธีการจำแนกได้ดังนี้

กำหนดให้ $f(x)$ เป็นฟังก์ชันการตัดสินใจและสมการทั่วไปของ SVM สามารถเขียนได้เป็น $f(x) = (w.\emptyset(x) + b)$, เมื่อ w เป็นเวกเตอร์ของค่าน้ำหนัก, b เป็นค่าอคติ (Bias) และทุกพารามิเตอร์จะเพิ่มค่า $\emptyset$ แทน Arbitrary Function ดังนั้นวัตถุประสงค์หลักของการจำแนกด้วย CNNSVM เพื่อทำการหาค่าเส้นแบ่งที่เหมาะสมที่สุดสำหรับข้อมูล $f(x)$ ดังนั้นกำหนดให้ข้อมูลการเรียนรู้เป็น $S = \{(x_i, y_i)\}_{i=1}^{m}$ เมื่อ $x_i \in R^n$ และ $y_i \in \{+1, -1\}$ สำหรับการจำแนกสองกลุ่ม และค่า Cost function มีสมการเป็น

$$minimize_w\{\lambda/2 \|w\|^2 + 1/m \sum_{i=1}^{m} l(x_i, y_i; w)\}, \quad (9)$$

เมื่อ $\lambda \geq 0$, และ $l(x_i, y_i; w) = \max\{0, 1 - y_i f(x_i)\}$. สำหรับเทอม $\max\{0, 1 - y_i f(x_i)\}$ จะถูกแทนด้วย Hinge-loss และกำหนดให้ $f(x_i)$ มีความสัมพันธ์กับ Kernel $K(\dots)$, และกำหนดให้ $\lambda/2 \|w\|^2$ ถูกใช้เป็นค่าสูงสุดของ Margin เมื่อเทอม $1/m \sum_{i=1}^{m} l(x_i, y_i; w)$, เป็นค่าต่ำสุดของค่าความคลาดเคลื่อนในการเรียนรู้และให้ ค่า C เป็นพารามิเตอร์ของการจำแนกคลาดเคลื่อน โดยที่มีความสัมพันธ์กับพารามิเตอร์ $\lambda = 1/mC$. โดยทั่วไปแล้วการเรียนรู้ในชุดของข้อมูลของ CNN และ SVM จะต้องทำให้ไปถึงค่าต่ำที่สุด $R(f(x_i), y_i)$ ดังนั้นในสมการที่ (9) จะถูกทำให้มีค่าต่ำสุดด้วยวิธีการ SGD [23] โดยแสดงขั้นตอนด้านล่าง

Algorithm 1: **Stochastic Sub-Gradient Projection**

Initialize: $\boldsymbol{w}$ such that $\|\boldsymbol{w}\| \leq (\boldsymbol{1}/\sqrt{\boldsymbol{\lambda}})$

For $\boldsymbol{t} = \{\boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{T}\}$ do

Swarm acquires $\boldsymbol{m}$ samples: $\{\boldsymbol{A_t(i)}|\boldsymbol{i} = \boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{m}\}$,

$\boldsymbol{A_t^+} = \{(\boldsymbol{X}, \boldsymbol{y}) \in \boldsymbol{A_t}: \boldsymbol{1} - \boldsymbol{y}(\boldsymbol{w_t}.\boldsymbol{X}) < \boldsymbol{1}\}$ and $\boldsymbol{w_{t+1}}$.

$$\boldsymbol{\eta_t} = \boldsymbol{1}/\sqrt{\boldsymbol{\lambda}}\left(\boldsymbol{w}_{\boldsymbol{t+\frac{1}{2}}}\right) = (\boldsymbol{1} - \boldsymbol{\eta_t}\boldsymbol{\lambda})\boldsymbol{w_t} + \frac{\boldsymbol{\eta_t}}{\boldsymbol{m}}\sum_{(\boldsymbol{x},\boldsymbol{y})\in\boldsymbol{A_t^+}} \boldsymbol{yX},$$

$$\boldsymbol{w_{t+1}} = \boldsymbol{minimize}\left\{\boldsymbol{1}, \frac{\boldsymbol{1}/\sqrt{\boldsymbol{\lambda}}}{\left\|\boldsymbol{w}_{\boldsymbol{t+\frac{1}{2}}}\right\|}\right\}\boldsymbol{w}_{\boldsymbol{t+\frac{1}{2}}},$$

End

Return $\boldsymbol{w}_{\boldsymbol{t+\frac{1}{2}}}$

รูปที่ 6 Algorithm Stochastic Sub-Gradient Projection

จากรูปที่ 6 Algorithm 1 จะมีข้อมูลเข้าพารามิเตอร์ m จากสมการ (9) และค่า T แทนจำนวนครั้งในการวนซ้ำ $\boldsymbol{t} = \{\boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{T}\}$ ในแต่ละครั้งของการวนซ้ำจะกำหนดค่าน้ำหนักเป็นเวกเตอร์ $\boldsymbol{w}$ ที่มีค่าบรรทัดฐานเป็น $\|\boldsymbol{w}\| \leq (\boldsymbol{1}/\sqrt{\boldsymbol{\lambda}})$ ในแต่ละครั้งของการวนซ้ำ $\boldsymbol{t}$ จะให้เซตของ $\{\boldsymbol{A_t(i)}|\boldsymbol{i} = \boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{m}\}$ ต้องผ่านกระบวนการ Swarm ในสมการที่ (9) ก่อนและสมการ

$$f(w; A_t) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{\boldsymbol{m}}\sum_{(x,y)\in A_t}^{m} l(x_i, y_i; w) \quad (10)$$

ตารางที่ 1 การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลภาพกับวิธีมาตรฐานด้วยข้อมูล Ground Truth [24]

Method	Objects Ground Truth																				
	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP (%)
GHM*	76.7	74.7	53.8	72.1	40.4	71.7	83.6	66.5	52.5	57.5	62.8	51.1	81.4	71.5	86.5	36.4	55.3	60.6	80.6	57.8	**64.7**
AGS*	82.2	83.0	58.4	76.1	56.4	77.5	88.8	69.1	62.2	61.8	64.2	51.3	85.4	80.2	91.1	48.1	61.7	67.7	86.3	70.9	**71.1**
NUS*	82.5	79.6	64.8	73.4	54.2	75.0	77.5	79.2	46.2	62.7	41.4	74.6	85.0	76.8	91.1	53.9	61.0	67.5	83.6	70.6	**70.5**
CNN	73.5	69.3	50.8	70.3	34.7	66.7	71.8	62.5	42.2	44.7	37.0	50.0	57.9	63.9	71.1	46.4	52.3	43.4	73.6	58.8	**57.0**
SVM	81.4	70.3	76.1	71.3	40.7	62.3	69.8	68.1	50.7	49.7	40.7	55.0	52.3	65.2	76.1	42.5	59.0	48.1	87.2	71.8	**61.9**
CNNSVM	86.6	79.9	75.3	78.5	48.4	75.8	83.5	78.0	55.1	61.6	55.7	68.0	78.9	75.9	86.5	49.5	67.3	61.6	87.1	70.3	**71.2**

6. ผลการทดลอง

การทดลองจำแนกภาพเพื่อให้เกิดประสิทธิภาพสูงสุดจะทำการทดลองเป็น 2 ส่วน (1) การเปรียบเทียบข้อมูลทั้งหมดจากฐานข้อมูลภาพทั่วไปที่เป็นมาตราฐานด้วย Ground Truth [12] ด้วยการใช้คำหลักเพียงอย่างเดียว ข้อมูลภาพทั้งหมดถูกคัดเลือกตามกลุ่มคำศัพท์หลักรวมทั้งหมด

4,800 ภาพแบ่งเป็น 16 กลุ่มคำศัพท์ กลุ่มละ 300 ภาพจากฐานข้อมูลภาพแอ็คชัน Pascal VOC 2007 [12] ข้อมูลภาพที่ใช้ในการทดลองทั้งหมดถูกคัดเลือกเฉพาะภาพที่เห็นวัตถุอย่างเด่นชัดและมีความสอดคล้องกับฐานข้อมูลพื้นฐานที่ประกอบด้วย 16 คำศัพท์หลัก ดังนี้ aero, bike, bird, boat, bus, car, cat, chair, cow, table, dog, horse, plant, sofa, train, และ television การทดลองเริ่มจากการเปรียบเทียบการจำแนกข้อมูลภาพตามกลุ่มภาพ Ground Truth จากฐานข้อมูล Pascal VOC 2007 สำหรับการทดสอบประสิทธิภาพการจำแนกข้อมูลภาพ โดยข้อมูลกลุ่มคำศัพท์ถูกคัดเลือกและมีพื้นฐานการทำลองเพื่อทำการเปรียบเทียบมาจากการทดลองของ J. Wu [24] ที่มีการคัดเลือกวัตถุบนภาพจาก 20 กลุ่ม และทำเป็นมาตราฐานของคำศัพท์ที่น่าเชื่อถือ ดังนั้นจากตารางที่ 1 จึงได้ใช้วัตถุและวิธีเปรียบเทียบพื้นฐานทั้งหมด 6 วิธีดังนี้ Generalized Hierarchical Matching (GHM) [24, 25], Ambiguity Guided Graph Shift (AGS) [24, 26], NUS [24, 27], Convolutional Neural Network (CNN), Support Vector Machines (SVM) และ Convolutional Neural Network - Support Vector Machines (CNNSVM) ดังแสดงผลลัพธ์ในตารางที่ 1 แสดงการจำแนกตามกลุ่มคำศัพท์จะเห็นว่าวิธี AGS ได้ค่าเฉลี่ยของความแม่นยำถึง 71.1% ซึ่งมีความใกล้เคียงกับวิธีการ CNNSVM ที่นำเสนอ 71.2% การจำแนกข้อมูลสำหรับการเรียนรู้ด้วย CNN ได้กำหนดอัตราการเรียนรู้ถูกกำหนดไว้ที่ 0.001 สำหรับ CNN และ 0.002 สำหรับ CNNSVM และ (2) การทดลองสำหรับเปรียบเทียบด้วยความสัมพันธ์จากการทำนาย ด้วยการวัดประสิทธิภาพของการทดลองการจำแนกข้อมูลด้วยความสัมพันธ์ระหว่างวัตถุและแอ็คชัน โดยแบ่งข้อมูลภาพจากฐานข้อมูล Ground Truth ข้างต้นเป็น 2 ชุด ชุดละ 2,800 ภาพ โดยที่ข้อมูลภาพใน Set I และ Set II ซึ่งมีการกำหนดค่า C เป็น 1 และจะต้องมีความสัมพันธ์ของวัตถุภายในภาพที่ประกอบด้วย แบบ implied-by, แบบ type-of และ แบบ mutually exclusive ดังแสดงผลการทดลองดังแสดงในตารางที่ 2 จะเห็นว่า การทดลองด้วยชุดข้อมูล Set I ประกอบด้วย วิธี Softmax [20, 21] ที่มีการกำหนดความสัมพันธ์ของแอ็คชันอยู่ที่ 20 แอ็คชัน และเพิ่มขึ้นเป็น 30 แอ็คชันสำหรับฐานข้อมูลชุด Set I ได้ค่าเฉลี่ย mAP 34.8% และ CNN SVM CNNSVM อยู่ที่ 48.2%, 53.3% และ 63.4% ตามลำดับ ส่วนวิธีการที่นำเสนอด้วยการจำแนกแบบ CNNSVM ที่มีการใช้ SGP ของชุดทดลอง Set I จะได้ค่าเฉลี่ย mAP อยู่ที่ 66.8% และชุดข้อมูล Set II จากตารางจะเห็นว่า CNNSVM-SGP สามารถได้ค่า mAP 71.4% และ 77.3% ของ 20, 30 แอ็คชันตามลำดับ ซึ่งเมื่อเฉลี่ย mAP ได้ 74.4% เมื่อเปรียบเทียบการจำแนกของ CNNSVM ของชุดข้อมูล Set I และ Set II ได้ค่าเฉลี่ย mAP 63.4% และ 68.4% จะแสดงให้เห็นว่าวิธีการ SGP สามารถเพิ่มประสิทธิภาพของการจำแนกได้สูงขึ้น แต่อย่างไรก็ตามเมื่อเทียบผลการทดลองของตารางที่ 1 ค่า CNNSVM ได้ค่า mAP สูงถึง 71.2%

ตารางที่ 2 ประสิทธิภาพของการจำแนกข้อมูลด้วยความสัมพันธ์ของแอ็คชัน

Method	Data SET I			Data SET II		
	20 actions	30 actions	Average mAP(%)	20 actions	30 actions	Average mAP(%)
Softmax	36.4	33.2	**34.8**	32.4	32.8	**32.6**
CNN	53.2	43.2	**48.2**	37.2	39.2	**38.2**
SVM	55.2	51.4	**53.3**	43.2	46.9	**45.1**
CNNSVM	62.3	64.4	**63.4**	67.3	69.5	**68.4**
CNNSVM-SGP	65.2	68.3	**66.8**	71.4	77.3	**74.4**
Average	**53.3**			**51.7**		

7. สรุปผลการทดลอง

งานวิจัยนี้ได้นำเสนอรูปแบบการจำแนกความหมายภาพด้วยการแทนข้อมูลภาพ จากความสัมพันธ์ของข้อมูลวัตถุภายในภาพด้วยคำศัพท์ที่มีความหมายเป็นแอ็คชั่นหรือการกระทำที่เกิดขึ้นบนภาพ โดยใช้อาศัยการเรียนรู้ความสัมพันธ์ของวัตถุและแอ็คชันที่เกิดขึ้น จากผลการทดลองจะเห็นว่าการใช้การทำนายความสัมพันธ์ของการแอ็คชัน ด้วยวิธีการจำแนกแบบ CNNSVM-SGD นั้นจะทำให้เกิดการจำแนกที่ดีกว่าการใช้แต่คำศัพท์เพียงอย่างเดียว แต่อย่างไรก็ตามความสัมพันธ์ที่กำหนดขึ้นมานั้นอาจยังไม่ครอบคลุม ทั้งหมดของฐานข้อมูลภาพ ดังนั้นอาจจำเป็นต้องมีการเพิ่มเติมแอ็คชั่นที่มีความเกี่ยวข้องและความสัมพันธ์ที่จำเป็นเพิ่มขึ้นเพื่อใช้สำหรับการทดลองต่อไป

References

[1] Liu GH, Yang JY, Li ZY. Content-based image retrieval using computational visual attention model. Pattern Recognition 2015;48:2554-66.

[2] Liu F, Xiang T, Hospedales MT, Yang W, Sun C. Semantic regularization for recurrent image annotation. 2017 IEEE Conference on Computer Vision and Pattern Recognition 2016;2872-80.

[3] Chinpanthana N. Similarity measure with directed universal hierarchical graph for digital image retrieval, Kasem Bundit Engineering Journal 2017;7(2):60-75. (In Thai)

[4] Zhou HT, Wang L, Ryu HK. Supporting keyword search for image retrieval with integration of probabilistic annotation 2015;7:6303-20.

[5] Chinpanthana N. A Study of feature extraction techniques used for content based image retrieval system. Christian University of Thai Journal, 2017;23:130-9. (In Thai)

[6] Jin J, Nakayama H. Annotation order matters: recurrent image annotator for arbitrary length image tagging. The 23rd international Conference on Pattern Recognition; 2016 Dec 4-8; Cancun, Mexico. p. 2452-7.

[7] Zhang H, Kyaw, Z, Chang SF, Chua TS. Visual translation embedding network for visual relation detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 5532-40.

[8] Lu C, Krishna R, Bernstein M, Li Fei-Fei. Visual relationship detection with language priors. European Conference on Computer Vision; 2016. p. 852–69.

[9] Patwardhan S, Pedersen T. Using wordNet based context vectors to estimate the semantic relatedness of concepts. Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together; 2006. p. 1-8.

[10] Deng J, Krause J, Berg CA, Li Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. IEEE Conference on Computer Vision and Pattern Recognition 2012.

[11] Chinpanthana N, Phiasai T, Deep textual searching for visual semantics of personal photo collections with a hybrid similarity measure. 2017 International Symposium on Computer Science and Intelligent Controls; 2017; Hungary.

[12] Everingham M, Gool VL, Williams CKI. Winn J, Zisserman A, The PASCAL visual object classes (VOC) challenge. International journal of computer vision 2010;88(2):303–38.

[13] Bangpeng Yao, Li Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2010. p. 9–16.

[14] Yao B., Li Fei-Fei, Modeling mutual context of object and human pose in human-object interaction activities. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2010. p. 17–24.

[15] Gupta A, Kembhavi A, Davis SL. Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 2009;31(10):1775–89.

[16] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE International Conference on Computer Vision (ICCV) 2015.

[17] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015.

[18] Ting S, Guohua G. Image retrieval method for deep neural network. International journal of Signal Processing, Image Processing and Pattern Recognition 2016;9(7):33-42.

[19] LeCun Y, Boser B, Denker SJ, Henderson D, Howard ER, Hubbard W, Jackel DL. Backpropagation applied to handwritten zip code recognition. Neural computation 1989;1(4):541-51.

[20] Socher R, Chen D, Manning DC, Ng YA. Reasoning with neural tensor networks for knowledge base completion. Advances in Neural Information Processing Systems 2013.

[21] Socher R, Karpathy A, Le VQ, Manning DC, Ng YA. Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistic 2013.

[22] Joachims T. Text categorization with support vector machines-learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning; 1998; Chemnitz, Germany. p. 137–42.

[23] Tsuruoka Y, Tsujii J, Ananiadou S. Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. Proceedings of the Asian Federation of Natural Language Processing 2009.

[24] Wu J, Yu Y, Huang C, Yu K. Deep multiple instance learning for image classification and auto-annotation. The IEEE Conference on Computer Vision and Pattern Recognition 2015.

[25] Chen Q, Song Z, Hua Y, Huang Z, and Yan S. Hierarchical matching with side information for image classification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012.

[26] Dong J, Xia W, Chen Q, Feng J, Huang Z, Yan S. Subcategory-aware object classification. IEEE Conference on Computer Vision and Pattern Recognition 2013.

[27] Song Z, Chen Q, Huang Z, Hua Y, Yan S. Contextualizing object detection and classification. IEEE Conference on Computer Vision and Pattern Recognition 2011.

ประวัติผู้เขียนบทความ

นัศพ์ชาณัณ ชินปัญช์ธนะ อาจารย์ประจำวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์ งานวิจัยทางด้านการประมวลผลภาพดิจิทัลทางด้านความหมายภาพ งานประมวลผลภาพประยุกต์กับงานทางด้านโรงงานอุตสาหกรรมการผลิต การจำแนกภาพดิจิทัล การรู้จำภาพดิจิทัล การประมวลผลภาพวิดีโอแบบเรียลไทม์ อีเมล์ nutchanun.cha@dpu.ac.th

Article History:

Received: June 8, 2018

Revised: April 25, 2019

Accepted: April 25, 2019