การศึกษาแบบจำลองความหมายเชิงประจักษ์สำหรับถอดรหัสรูปแบบการกระตุ้นของสมองบนเอฟเอ็มอาร์ไอ
A Study of Visual Semantic Models for Decoding fMRI activity patterns

ปิยะวัฒน์ แสงเพชร [1] ลือพล พิพานเมฆาภรณ์ [2] และสุวัจชัย กมลสันติโรจน์ [3]
[1] โปรแกรมวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสุราษฎร์ธานี, สุราษฎร์ธานี
[2,3] ภาควิชาวิทยาการคอมพิวเตอร์และสารสนเทศ คณะวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ
E-mail: piyawat@sru.ac.th[1], luepol.p@sci.kmutnb.ac.th[2] and suwatchai.k@sci.kmutnb.ac.th[3]

บทคัดย่อ

บทความวิจัยฉบับนี้นำเสนอการใช้ลักษณะเด่นเชิงประจักษ์ (visual feature) เพื่อสร้างแบบจำลองเชิงคำนวณสำหรับการทำนายคอนเซ็ปต์สุ่มเลือก (arbitrary concept) ซึ่งสัมพันธ์กับรูปแบบการกระตุ้นของสมองจากภาพเอฟเอ็มอาร์ไอ ในงานวิจัยนี้ผู้วิจัยตรวจสอบ 6 ชนิดของลักษณะเด่นเชิงประจักษ์ซึ่งได้จากการประมวลผลภาพระดับต่ำ ได้แก่ (1) color histogram, (2) color correlogram, (3) color moments, (4) edge direction histogram, (5) wavelet texture และ (6) Scale Invariance Feature Transform ซึ่งถูกใช้เพื่ออธิบายคอนเซ็ปต์ในรูปแบบของเวกเตอร์ความหมาย (semantic vector) ในการทำนายคอนเซ็ปต์ แบบจำลองซึ่งอธิบายความสัมพันธ์เชิงความหมายระหว่างจุดสมองและลักษณะเด่นเชิงประจักษ์เหล่านี้จะถูกสร้าง ผลการทดลองในการทำนายชุดคำนาม 60 คำจากชุดข้อมูลเอฟเอ็มอาร์ไอของมหาวิทยาลัยคอนเน็กกี้เมล่อนยืนยันว่าแบบจำลองความหมายเชิงประจักษ์ซึ่งสกัดจากภาพออนไลน์ที่ถูกแท็กด้วยคอนเซ็ปต์ที่เกี่ยวข้องในคลังภาพ Flickr สามารถทำนายคำหล่านี้ได้ไม่แตกต่างจากแบบจำลองความหมายเชิงภาษาซึ่งถูกยอมรับในปัจจุบัน

คำสำคัญ: การถอดรหัสสมอง ลักษณะเด่นเชิงประจักษ์ เอฟเอ็มอาร์ไอ การทำนายคอนเซ็ปต์ แบบจำลองเวกเตอร์สเปซ

Abstract

This study presents how visual features can be utilized for predicting arbitrary concepts from brain activity patterns. Six types of low-level image features, including (1) color histogram, (2) color correlogram, (3) edge direction histogram, (4) color moments, (5) wavelet texture and (6) Scale Invariance Feature Transform are extracted from online images associated with these concepts where (1)-(3) are based on global features and (4)-(6) are based on local ones. We also present a method of selecting images associated with each concept from Flickr, a large social image resource, based on tag relationship. Experimental results conducted on fMRI dataset provided by Carnegie Mellon University demonstrated that local-based visual feature models achieve encouraging performance of the task of predicting arbitrary words from fMRI activity patterns compared to global-based feature models and have no significant difference with state-of-the-art text-based feature models with manual user

Keywords: Brain decoding, visual features, fMRI, concept prediction and vector space model

1. บทนำ

ภาพวินิจฉัยระบบประสาท (neuroimaging) มักถูกนำมาใช้งานในทางการแพทย์เพื่อวินิจฉัยผู้ป่วยด้านระบบประสาทและสมอง หลายปีที่ผ่านมาเทคนิคจำนวนหนึ่งถูกพัฒนาสำหรับการสร้างภาพวินิจฉัยระบบประสาท เช่น Magnetoencephalography (MEG) และ Computed tomography (CT) เป็นต้น ระหว่างเทคนิคเหล่านี้ functional Magnetic Resonance Imaging (fMRI) ซึ่งวัดระดับการเปลี่ยนแปลงของออกซิเจนในกระแสเลือดที่ไหลไปตามเซลส์ประสาทในสมอง จากนั้นแปลงเป็นภาพถ่ายสมองสามมิติ (3D image) ถูกนำมาใช้เป็นเครื่องมือที่สำคัญในงานวิจัยด้านประสาทวิทยา (neuroscience), ปัญญาประดิษฐ์ (Artificial Intelligence) และการสร้างระบบเชื่อมต่อกับสมอง (Brain Computer Interface System) [16] ภาพเอฟเอ็มอาร์ไอโดยทั่วไปประกอบไปด้วยจุดสมอง (voxel) ประมาณ 20,000 จุด และจะถูกกระตุ้น (active) เมื่อสมองถูกใช้งาน เช่น การอ่าน การคิด และอารมณ์ เป็นต้น

งานวิจัยของ Tom Mitchell และคณะ [1] แสดงให้เห็นถึงความเป็นไปได้ในการทำนายรูปแบบการกระตุ้นจุดสมองจากข้อมูลเอฟเอ็มอาร์ไอซึ่งสัมพันธ์กับคอนเซ็ปต์ ผู้วิจัยนำเสนอแบบจำลองเชิงคำนวณซึ่งอธิบายความสัมพันธ์เชิงความหมายระหว่างจุดสมองและกลุ่มคำนาม ซึ่งผ่านกระบวนการเรียนรู้ของเครื่อง (machine learning) และมีบทบาทสำคัญอย่างยิ่งในการวิเคราะห์ภาพเอฟเอ็มอาร์ไอเพื่อจะอธิบายความสัมพันธ์ระหว่างจุดสมองและหน้าที่ส่วนต่างๆ ของสมอง [14-15] กับชุดข้อมูล ขณะที่ Palatucci และคณะ [2] นำเสนอแบบจำลองการทำนายคอนเซ็ปต์จากรูปแบบการกระตุ้นของจุดสมอง ประสิทธิภาพของแบบจำลองเหล่านี้ขึ้นอยู่กับวิธีที่ใช้ในการแสดงคอนเซ็ปต์หรือสเปซคอนเซ็ปต์ (concept space) ตัวอย่างเช่น ในงานวิจัย [1] ความสัมพันธ์เชิงความหมายระหว่างคำนามและคำกริยา 25 ตัวซึ่งถูกเลือกโดยนักภาษาศาสตร์จะถูกคำนวณเพื่อสร้างสเปซคอนเซ็ปต์ ขณะที่สเปซคอนเซ็ปต์จะถูกนิยามโดยเวกเตอร์ที่ได้จากการตอบคำถาม 218 ข้อที่เกี่ยวกับคุณสมบัติของแต่ละคอนเซ็ปต์ถึงแม้ว่าแบบจำลองเหล่านี้จะให้ความถูกต้องค่อนข้างสูง อย่างไรก็ตามข้อจำกัดของแบบจำลองเหล่านี้ก็คือไม่รองรับคอนเซ็ปต์ที่หลากหลาย

เพื่อที่จะแก้ข้อจำกัดดังกล่าว งานวิจัยจำนวนหนึ่งใช้เทคนิคประมวลผลภาษาธรรมชาติร่วมกับคลังข้อความขนาดใหญ่ เช่น Wikipedia [4], Google n-gram corpus [2] และ 50M web pages [3] เพื่อสกัดลักษณะเด่นเชิงภาษา (linguistic feature) ในการอธิบายคอนเซ็ปต์อัตโนมัติ อย่างไรก็ตามผลการทดลองพบว่าประสิทธิภาพของแบบ

จำลองเหล่านี้ยังคงให้ความถูกต้องไม่สูงมาก ปัญหาที่ยังท้าทายก็คือทำอย่างไรจึงจะได้รับลักษณะเด่นที่มีประสิทธิภาพสำหรับการอธิบายคอนเซ็ปต์ที่หลากหลาย (universal concept)

บทความวิจัยฉบับนี้ศึกษาการสำรวจลักษณะเด่นเชิงประจักษ์ (visual feature) ซึ่งได้จากวิเคราะห์ภาพ (image analysis) เพื่อใช้ในการอธิบายคอนเซ็ปต์แทนที่ลักษณะเด่นเชิงภาษา ข้อดีของการใช้ลักษณะเด่นเชิงประจักษ์ ได้แก่

- ง่ายที่จะถูกได้รับโดยการประมวลผลภาพระดับต่ำ (low level image processing)
- งานวิจัยไม่นานมานี้ยืนยันข้อดีในการใช้ลักษณะเด่นเชิงประจักษ์เพื่อแก้ปัญหาในงานประมวลผลภาษาธรรมชาติ [5][6]
- สามารถอธิบายคอนเซ็ปต์ในมุมมองที่ไม่สามารถถูกอธิบายโดยภาษา เช่น สี และพื้นผิว [6]

ผู้วิจัยตรวจสอบ 6 ชนิดของลักษณะเด่นเชิงประจักษ์ซึ่งใช้อธิบายเนื้อหาของภาพตาม 1) คุณลักษณะของสี (color) ได้แก่ color histogram, auto-correlogram และ color moments 2) คุณสมบัติของรูปร่าง (texture) ได้แก่ edge direction histogram, wavelet texture และ Scale Invariance Feature Transform (SIFT) ซึ่งใช้อธิบายคอนเซ็ปต์ในรูปของคำศัพท์ (lexical concept) เพื่อที่จะรองรับคอนเซ็ปต์ที่หลากหลาย ผู้วิจัยทำการคัดเลือกภาพที่เกี่ยวข้องกับคอนเซ็ปต์ซึ่งมาจากคลังภาพออนไลน์ Flickr[1] โดยปัจจุบัน Flickr มีภาพมากกว่า 14 ล้านซึ่งสัมพันธ์กับแท็ก (tag) ที่ถูกระบุโดยผู้ใช้งานอินเทอร์เน็ต ผู้วิจัยออกแบบการทดลองเพื่อเปรียบเทียบประสิทธิภาพในการทำนายคอนเซ็ปต์ซึ่งถูกอธิบายโดยลักษณะเด่นเชิงประจักษ์เหล่านี้จากข้อมูลเอฟเอ็มอาร์ไอของมหาวิทยาลัยคอนเน็กกี้เมล่อน [1] ผลการทดลองยืนยันว่าสเปซคอนเซ็ปต์ซึ่งสร้างขึ้นจากลักษณะเด่นเชิงประจักษ์เหล่านี้ สามารถทำนายคอนเซ็ปต์เชิงภาษาได้ไม่แตกต่างจากการใช้ลักษณะเด่นเชิงภาษาซึ่งถูกยอมรับในปัจจุบัน

2. Representation models for brain decoding

รูปที่ 1 แสดงแบบจำลองเชิงคำนวณเพื่อทำนายคอนเซ็ปต์ในรูปของคำนาม (noun) จากข้อมูลเอฟเอ็มอาร์ไอ ซึ่งถูกนำเสนอในงานวิจัย [2] จากรูปแบบจำลองประกอบไปด้วย 3 ส่วน ได้แก่ 1) ชั้นอินพุต (input layer) แทนด้วยเวกเตอร์ค่ากระตุ้นของจุดสมอง (voxel activation) $X=\{x_1, x_2, \ldots, x_n\}$ ซึ่งได้จากการประมวลผลภาพเอฟเอ็มอาร์ไอ 2) ชั้นกลาง (intermediate layer) แทนด้วยเวกเตอร์ลักษณะเด่นที่ใช้ในการอธิบายคอนเซ็ปต์ $Z=\{z_1, z_2, \ldots, z_k\}$ และ 3) ชั้นถอดรหัส (decoding layer) แทนด้วยเวกเตอร์ของคอนเซ็ปต์เป้าหมาย $W=\{w_1, w_2, \ldots, w_m\}$ โดยที่ความสัมพันธ์ระหว่างชั้นกลางและชั้นถอดรหัสถูกอธิบายโดยค่าลักษณะเด่นซึ่งสัมพันธ์กับแต่ละคอนเซ็ปต์ ขณะที่ความสัมพันธ์ระหว่าง เวกเตอร์ค่ากระตุ้นของจุดสมองและลักษณะเด่นสามารถถูกอธิบายแบบจำลองสมการถดถอย (1)

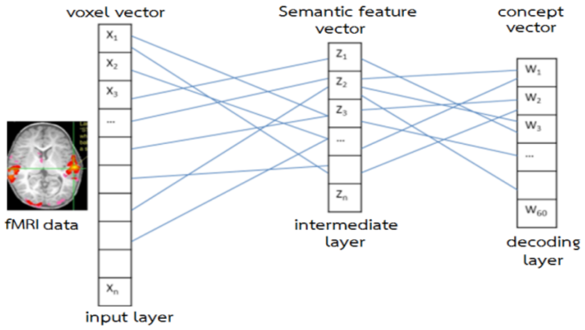
$$\vec{W}=\left(X^T X+\lambda I\right)^{-1} X^T Z \qquad (1)$$

โดยที่ $\vec{W}, I$ หมายถึงเมตริกซ์ค่าสัมประสิทธิ์และเมตริกซ์เอกลักษณ์

λ หมายถึงพารามิเตอร์ Regularization โดย λ สามารถจะถูกหาได้จากการทำ cross validation

เพื่อที่จะทำนายคอนเซ็ปต์ เริ่มต้นจากการประมวลผลภาพเอฟเอ็มอาร์ไอเพื่อแทนให้อยู่ในรูปเวกเตอร์ค่ากระตุ้นจุดสมอง X ค่ากระตุ้นของจุดสมองเหล่านี้จะถูกใช้เพื่อทำนายเวกเตอร์ของค่าลักษณะเด่น $\hat{Z}(X)$ จากนั้นเวกเตอร์ค่าลักษณะเด่นนี้จะถูกนำไปจัดลำดับความคล้ายคลึงระหว่างเวกเตอร์ของแต่ละคอนเซ็ปต์ $Z(w')$ โดยใช้ Cosine similarity และเวกเตอร์ของคอนเซ็ปต์ที่คล้ายคลึงมากที่สุดจะถูกเลือกเป็นคำตอบ

$$\underset{w' \in W}{argmax} \cos\left(\hat{Z}(X), Z(w')\right) \qquad (2)$$



ภาพที่ 1 แบบจำลองเชิงคำนวณสำหรับการทำนายคอนเซ็ปต์ [2]

เพื่อที่จะได้รับเวกเตอร์ Z สำหรับแต่ละคอนเซ็ปต์ งานวิจัย [2] นำเสนอเพื่อที่จะอธิบายคอนเซ็ปต์โดยใช้เวกเตอร์ 218 มิติ ที่ได้จากการตอบคำถามที่เกี่ยวกับคุณสมบัติของคอนเซ็ปต์ เช่น “Can you hold it?” และ “Is it animal?” เป็นต้น อย่างไรก็ตามเวกเตอร์ลักษณะเด่นนี้สามารถทำนายได้เฉพาะกลุ่มของคอนเซ็ปต์ที่ออกแบบไว้เท่านั้น และไม่สามารถขยายให้ครอบคลุมคอนเซ็ปต์มากขึ้น ต่อมางานวิจัยจำนวนหนึ่งนำเสนอเทคนิคการสกัดลักษณะเด่นเพื่ออธิบายคอนเซ็ปต์แบบอัตโนมัติ โดยอาศัยฐานความรู้ภาษาศาสตร์และคลังข้อความ เช่น งานวิจัย [4] ใช้ความสัมพันธ์เชิงความหมายระหว่างคอนเซ็ปต์ซึ่งได้จากเอกสาร Wikipedia ในการแสดงเวกเตอร์ Z หรืองานวิจัย [3] นิยามเวกเตอร์ Z ว่าเป็นเวกเตอร์การเกิดร่วม (co-occurrence) ระหว่างคอนเซ็ปต์กับคำอื่นๆ ที่พบในคลังข้อความภาษาอังกฤษ 50 ล้านเพจ ไม่ว่าอย่างไรก็ตามจากผลการทดลองพบว่าประสิทธิภาพการทำนายยังคงไม่สูงมากนักเมื่อเทียบกับการออกแบบโดยเวกเตอร์ลักษณะเด่นที่ถูกนิยามโดยการใช้นักภาษาศาสตร์

1 www.flickr.com

3. Visual semantic models

ในงานวิจัยฉบับนี้ ผู้วิจัยสำรวจ low-level features ที่สกัดได้จากการวิเคราะห์ภาพซึ่งสัมพันธ์กับแต่ละคอนเซ็ปต์ และใช้ลักษณะเด่นเชิงเหล่านี้เพื่อที่จะทำนายคอนเซ็ปต์ แนวคิดก็คือ low-level features เหล่านี้ง่ายที่จะถูกสกัดจากภาพและควรที่จะอธิบายคอนเซ็ปต์ในมุมมองที่แตกต่างจากการอธิบายด้วยคุณลักษณะเชิงภาษา ได้แก่ สี (color) พื้นผิว (texture) เป็นต้น ยิ่งไปกว่านั้นผลการทดลองจากงานวิจัยจำนวนหนึ่งแสดงให้เห็นถึงความสามารถของ low-level image features ในการจัดการคอนเซ็ปต์เชิงภาษา (lexical concept) [5][6]

3.1 Color-based features

สี (color) เป็นหนึ่งในสารสนเทศสำคัญซึ่งส่งผลต่อกระบวนการรับรู้และแปลความหมายของมนุษย์ (visual perception) ลักษณะเด่นเชิงประจักษ์ของสี (color feature) เป็นลักษณะเด่นพื้นฐานที่ถูกนำมาใช้ในการอธิบายเนื้อหาในรูปภาพ ในงานวิจัยนี้ ผู้วิจัยเลือกลักษณะเด่นเชิงประจักษ์ของสีในภาพซึ่งถูกใช้งานบ่อยๆ ได้แก่

- **ฮีสโตแกรมสี (Color histogram)** [9] ซึ่งจะอธิบายการกระจายของค่าสีสอดคล้องกับความถี่ เนื่องจากความหลากหลายของระบบสี ผู้วิจัยเลือกฮิสโตรแกรมสี LAB โดยที่ L คือค่าความสว่าง A คือค่าสีจากสัดส่วนของสีเขียวและสีแดง และ B คือค่าสีจากสัดส่วนของสีฟ้าและสีเหลือง เนื่องจากมีความทนทานต่อสัญญาณรบกวน เพื่อที่จะลดมิติของฮิสโตรแกรมสี ผู้วิจัยแบ่งช่วงค่าในแต่ละองค์ประกอบ LAB ออกเป็น 4 ช่วงเท่าๆ กัน และนับความถี่ในแต่ละช่วง ดังนั้นฮิสโตรแกรมสีที่ใช้ในงานวิจัยนี้มีขนาด 64 มิติ (4x4x4)
- **ความสัมพันธ์เชิงพื้นที่ระหว่างคู่สี (Color correlogram)** กำหนดให้ p_i และ p_j เป็นสองพิกเซลใดๆ ในภาพ ซึ่งมีระยะห่างเท่ากับ $|p_i - p_j| = k$ ความสัมพันธ์เชิงพื้นที่ระหว่างคู่สี C_x และ C_y ถูกนิยามว่าเป็นความน่าจะเป็นที่จะเกิดสี C_x ในพิกเซล p_i ที่มีระยะห่าง k จากพิกเซล p_j ที่พบสี C_y ในบทความนี้เลือกใช้ color correlogram HSV ซึ่งถูกนำเสนอใน [10] โดยจะแบ่งช่วงค่าของสีทั้งหมด 36 ช่วง และเซตของระยะทาง $k = \{1,3,5,7\}$ ดังนั้นเวกเตอร์คุณลักษณะเด่น color correlogram จะเท่ากับ 36x4 = 144 มิติ
- **Color moments** [9] ซึ่งจะอธิบายเนื้อหาของภาพด้วยค่าทางสถิติ 3 ค่า ได้แก่ ค่าเฉลี่ย (mean) ค่าความแปรปรวน (variance) และค่าความเบ้ (skewness) ของแต่ละองค์ประกอบสี โดยทั่วไป color moments จะให้เวกเตอร์ขนาด (3x3x3) 9 มิติ ซึ่งอาจไม่เพียงพอในการแยกแยะภาพ ในงานวิจัยนี้จะสร้างกริด (grid) บนภาพ ขนาด 5 x 5 และคำนวณค่า color moments ในแต่ละกริด ซึ่งจะทำให้เวกเตอร์มีขนาด 5x5x9 = 225 มิติ

3.2 Texture-based features

ในการประมวลผลภาพถ่ายดิจิตอล เนื้อภาพ (texture) ถูกนิยามว่าเป็นโครงสร้างย่อยที่ปรากฏในเอกสารภาพ ลักษณะเด่นของเนื้อภาพจะอธิบายเนื้อหาของภาพโดยการสำรวจโครงสร้างย่อยที่ปรากฏในภาพซึ่งจะมีความทนทานต่อสัญญาณรบกวนและให้รายละเอียดของภาพได้ดีกว่าการใช้ลักษณะเด่นของสี ในบทความวิจัยฉบับนี้ผู้วิจัยเลือกที่จะสำรวจลักษณะเด่นเชิงประจักษ์ที่เกี่ยวของพื้นผิวซึ่งถูกใช้งานบ่อย ประกอบไปด้วย

- **ฮีสโตรแกรมของทิศทางขอบภาพ (Edge direction histogram)** [11] ซึ่งจะอธิบายการกระจายตัวของทิศทางขอบภาพที่ปรากฏในรูปถ่าย โดยทั่วไปเวกเตอร์ของฮีสโตรแกรมของทิศทางขอบภาพจะมีขนาด 73 มิติ โดย 72 มิติแรกจะแสดงความถี่ของการเกิดขอบภาพในทิศทางซึ่งถูกแบ่งทีละ 5 องศา ส่วนมิติสุดท้ายจะแสดงจำนวนพิกเซลที่ไม่เกิดขอบภาพ ในงานวิจัยนี้ใช้เทคนิคการหาขอบภาพ Canny ร่วมกับโอเปอเรเตอร์ Sobel เพื่อที่จะหาทิศทางของขอบภาพโดยการคำนวณค่าเกรเดียน (gradient) ของแต่ละจุด
- **เนื้อภาพของเวฟเล็ต (Wavelet texture)** [12] การแปลงเวฟเล็ตปัจจุบันเป็นเทคนิคที่ได้รับความนิยมสำหรับการวิเคราะห์สัญญาณ ในการแปลงเวฟเล็ตของภาพเอกสารจะใช้การแปลงเวฟเล็ตแบบดีสครีท 2 มิติ โดยการแยกองค์ประกอบของภาพ (decomposition) เป็นแบนด์ย่อยๆ หลายระดับ โดยที่แต่ละระดับภาพต้นฉบับจะถูกแบ่งออกเป็นแบนด์ย่อยของความถี่จำนวน 4 แบนด์ ถูกนิยามว่าเป็น **LL, LH, HL** และ **HH** โดยที่ **L** หมายถึงความถี่ต่ำ (low frequency) และ **H** หมายถึงความถี่สูง (high frequency) หลังจากนั้น 2 ชนิดของการแปลงเวฟเล็ต ได้แก่ Pyramid-structured Wavelet Transform (PWT) ซึ่งจะถูกใช้เพื่อที่จะแยกองค์ประกอบของแบนด์ย่อย **LL** และ Tree-structured wavelet transform (TWT) จะถูกใช้ เพื่อที่จะแยกองค์ประกอบของแบนด์ย่อย **LH, HL** และ **HH** ที่แต่ละระดับ หลังจากขั้นตอนการแยกองค์ประกอบเวกเตอร์ลักษณะเด่นซึ่งถูกคำนวณจากค่าเฉลี่ย (mean) และค่าเบี่ยงเบนมาตรฐาน (standard deviation) ของค่าสัมประสิทธิ์เวฟเล็ตที่ได้ในแต่ละแบนด์ย่อยและแต่ละระดับ ในงานวิจัยนี้กำหนดระดับความลึกเท่ากับ 3 ในการแยกองค์ประกอบของภาพต้นฉบับ ดังนั้น PWT จะเป็นผลลัพธ์ในการสกัดลักษณะเด่นจำนวน 24 (3x4x2) ค่า และ TWT จะเป็นผลลัพธ์ในการสกัดลักษณะเด่นจำนวน 104 (52x 2) ค่า ซึ่งเมื่อนำมารวมกันจะทำให้เวกเตอร์ของเวฟเล็ตมีจำนวนเท่ากับ 128 มิติ
- **Bag-of-Words of Scale Invariance Feature Transform (SIFT)** [13] เป็นลักษณะเด่นเชิงประจักษ์สำหรับการวิเคราะห์เนื้อภาพที่ได้รับความสนใจจากนักวิจัยจำนวนมาก เนื่องจากคุณสมบัติของลักษณะเด่น SIFT ซึ่งทนทานต่อการเปลี่ยนแปลงของภาพ เช่น การเลื่อน (translation) การหมุน (orientation) การเปลี่ยนขนาด (scaling) และการเปลี่ยนระดับแสง (illumination) การสกัดลักษณะเด่น SIFT จะประกอบไปด้วย 2 ขั้นตอน ได้แก่ 1) การหาความแตกต่างของภาพระดับสีเทาด้วยตัวกรองแบบเกาส์เซี่ยน (Gaussian filter) ที่ความละเอียดต่างๆ กัน เพื่อหาตำแหน่งของ

keypoint ในภาพ 2) การสกัดลักษณะเด่นของแต่ละ keypoint ซึ่งอธิบายเนื้อภาพในบริเวณรอบๆ keypoint ดังกล่าว โดยทั่วไปลักษณะเด่นของ keypoint จะแทนในรูปของเวกเตอร์ขนาด 128 มิติ อย่างไรก็ตามจำนวน keypoint ในภาพอาจมีมากและหลายๆ keypoint มีลักษณะเด่นที่คล้ายคลึงกัน จึงมีการแบ่งกลุ่ม keypoint โดยใช้เทคนิค k-mean และนำตัวแทนแต่ละกลุ่มสร้างเป็นเวกเตอร์ขนาด k มิติ เรียกว่าเวกเตอร์ถุงคำ (bag-of-words) ในงานวิจัยนี้กำหนดค่า k เท่ากับ 500 เพื่อแบ่งกลุ่ม keypoint ในภาพ ดังนั้นเวกเตอร์ลักษณะเด่น SIFT จะมีขนาด 500 มิติ

3.3 Image-based semantic vectors

เพื่อที่จะสกัดเวกเตอร์ของภาพในการอธิบายคอนเซ็ปต์ ผู้วิจัยเลือกที่จะใช้ภาพจากชุดข้อมูล NUS-WIDE [7] ซึ่งประกอบไปด้วยภาพจำนวน 269,648 ภาพที่ถูกรวบรวมจากคลังภาพออนไลน์ Flickr แต่ละภาพในชุดข้อมูล NUS-WIDE จะสัมพันธ์กับแท็กซ์ (tag) ซึ่งถูกใช้เพื่ออธิบายเนื้อหาของภาพ โดยจะมีจำนวนแท็กซ์ทั้งสิ้น 5,018 แท็กซ์ ซึ่งถูกใช้มากกว่า 100 ครั้ง ขณะที่ค่าเฉลี่ย (average) ของจำนวนแท็กซ์ต่อภาพใน NUS-WIDE จะอยู่ที่ 8.5 แท็กซ์

กำหนดให้ $W=\{w_1,w_2,\ldots,w_m\}$ เป็นเซตของ m คอนเซ็ปต์ ผู้วิจัยเลือกที่จะวิเคราะห์ภาพจากชุดข้อมูล NUS-WIDE เฉพาะที่ถูกแท็กซ์ตั้งแต่ 4 คำขึ้นไป โดยจะต้องมีอย่างน้อย 1 แท็กซ์ซึ่งตรงกับคอนเซ็ปต์ อย่างไรก็ตามพบว่าจำนวนภาพที่จะใช้ในการสกัดลักษณะเด่นเชิงประจักษ์ยังคงมีมากและแต่ละคอนเซ็ปต์ก็มีจำนวนภาพที่แตกต่างกัน ผู้วิจัยนำเสนอวิธีการเลือกภาพสำหรับแต่ละคอนเซ็ปต์โดยการวิเคราะห์ความสัมพันธ์ระหว่างแท็กซ์และคอนเซ็ปต์ แนวคิดก็คือว่าภาพซึ่งถูกอธิบายด้วยแท็กซ์ซึ่งมีความสัมพันธ์กับคอนเซ็ปต์บ่อยๆ น่าที่จะมีแนวโน้มที่จะเกี่ยวข้องกับคอนเซ็ปต์มากกว่าภาพซึ่งถูกอธิบายด้วยแท็กซ์ทั่วๆ ไป

กำหนดให้ $T=\{t_1,t_2,\ldots,t_n\}$ เป็นเซตของแท็กซ์ซึ่งมีความถี่ (frequency) มากที่สุด n อันดับแรกในชุดข้อมูล NUS-WIDE (ในงานวิจัยนี้ผู้วิจัยกำหนด $n=1{,}000$) ค่าความสัมพันธ์ระหว่างแท็กซ์ t_i คอนเซ็ปต์ w_j แสดงได้ในสมการ (2)

$$r_{ij}=\frac{f_{ij}}{f_i+f_j-f_{ij}} \tag{2}$$

โดยที่ f_{ij} หมายถึงความถี่ซึ่งแท็กซ t_i และคอนเซ็ปต์ w_j ปรากฏในภาพเดียวกัน โดยที่ค่า r_{ij} จะอยู่ระหว่าง 0 ถึง 1

หลังจากค่าความสัมพันธ์ระหว่างแท็กซ์และคอนเซ็ปต์ถูกคำนวณแล้ว แต่ละภาพ I จะถูกนำไปคำนวณค่าระดับคะแนน (score) เพื่อใช้ในการเลือกภาพสำหรับแต่ละครเซ็ปต์โดยใช้สมการที่ (3)

$$S(w_j,I)=\sum_{t_i\in T(I)} r_{ij} \tag{3}$$

โดยที่ $S(w_j,I)$ หมายถึงค่าระดับคะแนนของภาพ I สำหรับคอนเซ็ปต์ w_j $T(I)$ หมายถึงเซตของแท็กซ์ 1k ซึ่งปรากฏในภาพ I สุดท้ายภาพที่ได้รับค่าระดับคะแนนสูงที่สุด k ลำดับแรกในแต่ละคอนเซ็ปต์จะถูกเลือกเพื่อใช้ในการสกัดเวกเตอร์ลักษณะเด่นเชิงประจักษ์ ในงานวิจัยนี้เลือก $k=100$ เนื่องจากมีแนวโน้มที่จะให้ผลลัพธ์ที่ดีที่สุดสำหรับชุดข้อมูลทดสอบ

ในขั้นตอนสุดท้ายเวกเตอร์ลักษณะเด่นเชิงประจักษ์ของภาพซึ่งถูกเลือกสำหรับแต่ละคอนเซ็ปต์จะถูกนำมาคำนวณค่าเฉลี่ย (mean) และทำการนอมัลไลต์ (normalize) ค่าเหล่านี้โดยใช้ z-score ก่อนนำไปสร้างแบบจำลองเพื่อถอดรหัสสมองเหมือนกับในหัวข้อ 2

4. การทดลอง (Experiment)

ในหัวข้อนี้ผู้วิจัยอธิบายรายละเอียดของชุดข้อมูล การออกแบบทดลองและผลลัพธ์

4.1 ชุดข้อมูล fMRI

งานวิจัยนี้ใช้ชุดข้อมูลเอฟเอ็มอาร์ไอจากมหาวิทยาลัยคอนเน็กกี้เมล่อน[2] ซึ่งประกอบไปด้วยข้อมูลเอฟเอ็มอาร์ไอจากอาสาสมัคร 9 คน (P1-P9) โดยอาสาสมัครแต่ละคนจะถูกให้ดูชุดภาพลายเส้นจำนวน 60 ภาพซึ่งจะแทนด้วยคอนเซ็ปต์ในรูปคำนาม (noun) สำหรับแต่ละภาพอาสาสมัครจะถูกบันทึกภาพสมองเอฟเอ็มอาร์ไอจำนวน 5 รอบ ดังนั้นแต่ละคนจะถูกบันทึกภาพเอฟเอ็มอาร์ไอจำนวน 5x60=360 ภาพ และทำการเฉลี่ยรวมให้เหลือเพียง 60 ภาพสำหรับแต่ละคอนเซ็ปต์ ในแต่ละภาพเอฟเอ็มอาร์ไอจะถูกแปลงให้อยู่ในรูปของเวกเตอร์ค่ากระตุ้นของจุดสมอง (voxel) ประมาณ 20,000 จุดโดยเฉลี่ย ซึ่งจะถูกลดมิติลงก่อนที่จะนำไปใช้ในการสร้างแบบจำลองเพื่อทำนายคอนเซ็ปต์จากเอฟเอ็มอาร์ไอ

ในงานวิจัย [1] และ [2] เวกเตอร์เอฟเอ็มอาร์ไอจะถูกลดมิติให้เหลือ 500 มิติ ดังนั้นในการทดลองนี้ ผู้วิจัยจึงเลือกใช้ค่านี้สำหรับการสร้างแบบจำลอง ยิ่งไปกว่านั้นในงานวิจัยนี้ ผู้วิจัยใช้เทคนิค searchlight ซึ่งถูกนำเสนอในงานวิจัย [8] เพื่อที่จะลดมิติเวกเตอร์เอฟเอ็มอาร์ไอซึ่งสัมพันธ์กับ 60 คอนเซ็ปต์

4.2 text-based semantic vectors

งานวิจัยนี้เลือกที่จะเปรียบเทียบประสิทธิภาพของเวกเตอร์เชิงความหมายซึ่งสร้างจากลักษณะเด่นเชิงประจักษ์จากเอกสารภาพกับสองแบบจำลองซึ่งถูกสร้างจากเวกเตอร์ลักษณะเด่นเชิงภาษา ได้แก่

- **Verb25** ถูกนำเสนอในงานวิจัย [1] โดยที่แต่ละคอนเซ็ปต์ในรูปของคำนามจะถูกอธิบายโดยเวกเตอร์การเกิดร่วม (co-occurrence) ระหว่างคำนามและคำกริยา (verb) จำนวน 25 คำ เช่น “see”, “hear”, “taste”, “smell”, และ “eat” เป็นต้น คำกริยาเหล่านี้เป็นกริยาทั่วไปที่มักจะเกิดร่วมกับคำนามเชิงวัตถุ (concrete noun) ในข้อความภาษาอังกฤษ และถูกออกแบบโดยนักภาษาศาสตร์ จากผลการทดลองพบว่าเวกเตอร์ Verb25 จะให้ผลลัพธ์ที่น่าพอใจบนชุดข้อมูลทดลองนี้
- **Human218** ถูกนำเสนอในงานวิจัย [2] เวกเตอร์ human218 จะอธิบายคอนเซ็ปต์ในรูปของเวกเตอร์ขนาด 218 มิติที่ได้จากการตอบ

2 http://www.cs.cmu.edu/afs/cs/project/theo37/www/science2008/data.html

คำถามทั้งหมด 218 คำถามที่เกี่ยวข้องกับคุณสมบัติของแต่ละคอนเซ็ปต์ เช่น “Can you hold it?” Is it a manmade? และ “Is it animal?” เป็นต้น คำถามเหล่านี้ถูกออกแบบโดยนักภาษาศาสตร์ เพื่อที่ได้รับเวกเตอร์ human218 ผู้วิจัยรวบรวมผลการตอบคำถามจากผู้ใช้งานอินเทอร์เน็ตและทำการหาค่าเฉลี่ยในแต่ละคำถามสัมพันธ์กับแต่ละคอนเซ็ปต์ สอดคล้องกับผลการทดลองในงานวิจัย [3] ซึ่งจะทำการเปรียบเทียบลักษณะเด่นเชิงภาษาสำหรับการทำนายคอนเซ็ปต์จากเอฟเอ็มอาร์ไอพบว่าเวกเตอร์ human218 ให้ผลลัพธ์ดีที่สุดบนชุดข้อมูลทดลองนี้เปรียบเทียบกับลักษณะเด่นเชิงภาษาอื่น

4.3 การวัดประสิทธิภาพของแบบจำลอง

จากที่กล่าวมาข้างต้น ความสามารถในการอธิบายคอนเซ็ปต์ที่หลากหลายคือคุณสมบัติสำคัญสำหรับเวกเตอร์เชิงความหมาย เพื่อที่จะวัดความสามารถดังกล่าว งานวิจัย [2] นำเสนอวิธีการ Leave-two-out cross validation โดยการ train แบบจำลองของอาสาสมัครแต่ละคนด้วยภาพเอฟเอ็มอาร์ไอจำนวน 58 ภาพ (1 ภาพสำหรับ 1 คอนเซ็ปต์) หลังจากนั้นทำการทดสอบประสิทธิภาพของแบบจำลองโดยนำภาพเอฟเอ็มอาร์ไอ 2 ภาพที่ถูกไม่ได้ถูก train ไปทำนายค่าของเวกเตอร์ลักษณะเด่น ผลการทำนายค่าดังกล่าวจะถูกนำไปเปรียบเทียบกับเวกเตอร์ลักษณะเด่นของแต่ละคอนเซ็ปต์และเลือกคอนเซ็ปต์ที่มีค่าเวกเตอร์ใกล้เคียงมากที่สุดเป็นคำตอบ กระบวนการนี้จะถูกดำเนินการต่อไปเรื่อยๆ โดยสับเปลี่ยนหมุนเวียนเพื่อทำนายทุกๆ คู่ของคอนเซ็ปต์ ดังนั้นแบบจำลองจะถูก train ทั้งหมด $\binom{60}{2}=1{,}770$ รอบ สำหรับ 60 คอนเซ็ปต์ ซึ่งจะนำไปสู่การเปรียบเทียบผลการทำนายทั้งสิ้น 1,770 รอบ x 2 คอนเซ็ปต์ = 3,540 ครั้ง ดังนั้นความแม่นยำ (accuracy) ในการทำนายคอนเซ็ปต์สามารถที่จะถูกนิยามว่าเป็นสัดส่วนของจำนวนครั้งในการทำนายคอนเซ็ปต์ที่ถูกต้อง (corrected prediction) และจำนวนครั้งในการทำนายทั้งหมด (total prediction) ในการทดลองนี้ผู้วิจัยเลือกที่จะใช้วิธี leave-two-out cross validation ในการเปรียบเทียบประสิทธิภาพของเวกเตอร์ลักษณะเด่น

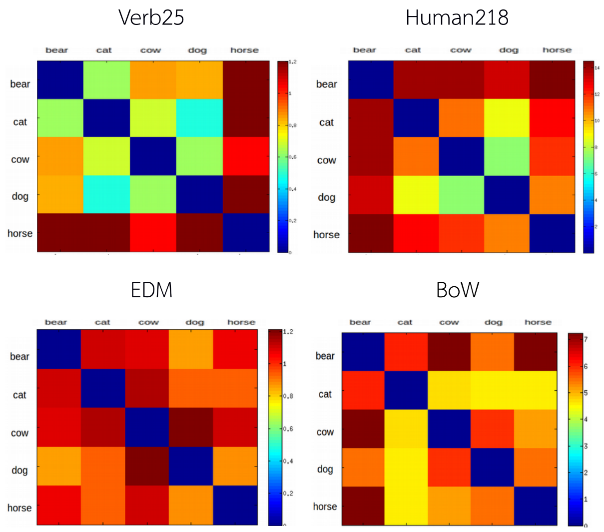
4.4 ผลการทดลอง

ในงานวิจัยนี้ 6 เวกเตอร์ลักษณะเด่นเชิงประจักษ์ ซึ่งประกอบไปด้วย color histrogram (CH), color correlogram (CORR), color moment (CM), wavelet texture (WT), edge direction histogram (EDH) และ Bag-of-word of SIFT (BoW) จะถูกเปรียบเทียบกับ Verb25 และ Human218 ซึ่งเป็นเวกเตอร์ลักษณะเด่นที่มาจาก text ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองซึ่งสร้างขึ้นจากเวกเตอร์ลักษณะเด่นด้วยวิธีต่างๆ แสดงได้ดังตารางที่ 1 จะเห็นได้ว่า BoW ซึ่งอยู่ในกลุ่มของลักษณะเด่นเชิงประจักษ์ที่เกี่ยวของพื้นผิวให้ความแม่นยำในการทำนาย 60 คอนเซ็ปต์ สำหรับอาสาสมัคร P1 ถึง P9 สูงที่สุด เมื่อเปรียบเทียบกับลักษณะเด่นเชิงประจักษ์แบบอื่นๆ ตามด้วย Human218 และ Verb25 ซึ่งสร้างจากลักษณะเด่นเชิงภาษา [1-3] ผลลัพธ์นี้แสดงให้เห็นถึงความสามารถของ BoW ในการอธิบายคอนเซ็ปต์ที่มีความใกล้เคียงกับการอธิบายเนื้อหาในเอกสารข้อความ (text representation) แสดงให้เห็นว่าสามารถนำคุณลักษณะเด่นเชิงประจักษ์มาใช้ในการทำนายแทนคุณลักษณะเด่นเชิงภาษาได้ [7] ยิ่งไปกว่านั้นลักษณะเด่นเชิงประจักษ์ที่มาจากเนื้อภาพ (texture) ซึ่งโดยทั่วไปจะมีความทนทานต่อการเปลี่ยนแปลงของภาพได้ดี มีแนวโน้มที่จะให้ความแม่นยำในการทำนายคอนเซ็ปต์ที่สูงกว่าลักษณะเด่นเชิงประจักษ์ที่มาจากสี (color) ยกเว้น WT ซึ่งให้ความแม่นยำในการทำนายต่ำที่สุด

เพื่อที่จะแสดงความสามารถของเวกเตอร์เชิงความหมายในการแยกแยะคอนเซ็ปต์ที่มีความใกล้เคียงกัน ผู้วิจัยเลือกเวกเตอร์ซึ่งถูกสร้างโดยลักษณะเด่นเชิงประจักษ์ที่ให้ความแม่นยำ 2 อันดับแรก ได้แก่ BoW และ EDM เปรียบเทียบกับ Verb25 และ Human 218 ในการแยกความแตกต่างระหว่างคอนเซ็ปต์ในกลุ่ม animal ซึ่งประกอบไปด้วย “bear”, “cat”, “cow”, “dog” และ “horse” โดยวัดค่าความคล้ายคลึงระหว่างเวกเตอร์ลักษณะเด่นซึ่งอธิบายคอนเซ็ปต์เหล่านี้ด้วยวิธี Cosine similarity สอดคล้องกับภาพที่ 2 BoW มีแนวโน้มที่จะแยกความแตกต่างระหว่างคอนเซ็ปต์ที่ใกล้เคียงกันได้ดีพอๆ กับ Human218

ตารางที่ 1 ผลการเปรียบเทียบความแม่นยำ (%) ของแบบจำลองที่สร้างโดยเวกเตอร์ลักษณะเด่นแบบต่างๆ

Category	Method	P1	P2	P3	P4	P5	P6	P7	P8	P9	Mean Rank	Rank
Color-based feature	CH	44.69	42.03	41.64	38.22	38.19	28.31	37.32	30.71	43.14	4.78	5
	CORR	42.60	41.05	36.58	36.05	37.68	26.30	34.58	34.04	39.01	5.89	6
	CM	27.26	23.50	24.89	29.15	21.69	17.99	25.54	18.42	23.84	7.00	7
Texture-based feature	WT	20.37	17.74	17.71	16.84	16.75	15.56	16.53	15.88	17.37	8.00	8
	EDH	43.87	47.15	40.45	45.65	42.06	29.41	36.38	36.13	45.11	4.11	4
	BoW	**57.26**	49.58	**54.60**	53.67	52.12	49.15	44.80	**60.59**	**51.67**	1.89	**1**
Text-based features	Verb25	56.30	53.19	53.14	**54.58**	**53.28**	42.94	47.97	46.05	45.00	2.22	**3**
	Human218	50.40	**55.37**	46.69	43.98	52.40	**49.41**	**60.96**	46.53	46.53	2.11	**2**



ภาพที่ 2 ค่าความสัมพันธ์ระหว่างคอนเซ็ปต์ในกลุ่ม animal

7. สรุปผลการทดลอง

บทความวิจัยฉบับนี้ทำการศึกษาการใช้ลักษณะเด่นเชิงประจักษ์ซึ่งได้จากการวิเคราะห์ภาพในการสร้างแบบจำลองเพื่อทำนายคอนเซ็ปต์เชิงวัตถุจากภาพเอฟเอ็มอาร์ไอ ผู้วิจัยทำการเลือกลักษณะเด่นเชิงประจักษ์ 6 ชนิด ได้แก่ (1) color histogram, (2) color correlogram, (3) color moments (4) edge direction histogram, (5) wavelet texture และ (6) Bag-of-Words (BoW) of SIFT โดยที่ (1)-(3) เป็นลักษณะเด่นที่มาจากสี ขณะที่ (4)-(5) เป็นลักษณะเด่นที่มาจากเนื้อภาพ ซึ่งส่งผลกระทบต่อกระบวนการรับรู้และแปลความหมายของสมองมนุษย์ ผู้วิจัยทำการออกแบบการทดลองและเปรียบเทียบผลลัพธ์ของแบบจำลองซึ่งสร้างขึ้นจากลักษณะเด่นเชิงประจักษ์เหล่านี้กับแบบจำลองซึ่งสร้างจากลักษณะเด่นเชิงภาษา human218 และ Verb25 ซึ่งให้ผลลัพธ์ที่ดีที่สุดในปัจจุบัน ผลการทดลองบนชุดข้อมูลเอฟเอ็มอาร์ไอจากมหาวิทยาลัยคอนเน็กกี้เมล่อนพบว่า BoW ให้ผลลัพธ์ได้ดีที่สุด

8. เอกสารอ้างอิง

[1] T. Mitchell, S. V. Shinkareva, A. Carlson, K. Chang, V. L. Malave, R. A. Mason, and M. A. Just. “Predicting human brain activity associated with the meanings of nouns”, Science, 320(5880):1191–5, 2008.

[2] Palatucci, Mark, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell. "Zero-shot learning with semantic output codes." In Advances in neural information processing systems, pp. 1410-1418. 2009.

[3] Murphy, B., Talukdar, P., & Mitchell, T.. Selecting corpus-semantic models for neurolinguistic decoding. In Proceedings of the Sixth International Workshop on Semantic Evaluation pp. 114-123, 2012.

[4] Luepol, P., Ludmilla, T., Guigue, V., and Artières, T. Designing semantic feature spaces for brain-reading. In European Symposium on Artificial Neural Networks, 2015.

[5] Bergsma, S., & Goebel, R. Using Visual Information to Predict Lexical Preference. In RANLP pp. 399-405 2011.

[6] Cai, H., Huang, Z., Zhu, X., Zhang, Q. and Li, X., Multi-Output Regression with Tag Correlation Analysis for Effective Image Tagging. In International Conference on Database Systems for Advanced Applications, pp. 31-46, 2014.

[7] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. "NUS-WIDE: A Real-World Web Image Database from National University of Singapore", ACM International Conference on Image and Video Retrieval. Greece. Jul. 8-10, 2009.

[8] ปริญญา เจริญวรเกียรติ ลือพล พิพานเมฆาภรณ์ และ สุวัจชัย กมลสันติโรจน์. การสร้างสเปซคุณลักษณะเด่นสำหรับการวิเคราะห์ข้อมูลเอฟ-เอ็มอาร์ไอ. ใน 19th International conference of Computer Science and Engineering (ICSEC) (Thai Track) ปี พ.ศ. 2558.

[9] L. G. Shapiro and G. C. Stockman. Computer Vision. Prentice Hall, 2003.

[10] J. Huang, S. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlogram. In IEEE Conf. on Computer Vision and Pattern Recognition, pages 762–768, June 1997.

[11] D. K. Park, Y. S. Jeon, and C. S. Won. Efficient use of local edge histogram descriptor. In ACM Multimedia, 2000.

[12] B. S. Manjunath and W.-Y. Ma. Texture features for browsing and retrieval of image data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(8):837–842, August 1996.

[13] Sivic, Josef, and Andrew Zisserman. "Video Google: A text retrieval approach to object matching in videos.", In Proceedings. Ninth IEEE International Conference on Computer Vision, pp. 1470-1477, 2003.

[14] Norman, Kenneth A., et al. "Beyond mind-reading: multi-voxel pattern analysis of fMRI data." Trends in cognitive sciences 10.9 (2006): 424-430.

[15] Carroll, Melissa K., et al. "Prediction and interpretation of distributed neural activity with sparse models." NeuroImage 44.1 (2009): 112-122.

[16] Daly, Janis J., and Jane E. Huggins. "Brain-Computer Interface: Current and Emerging Rehabilitation Applications." Archives of physical medicine and rehabilitation 96.3 (2015): S1-S7.