**OBJECT DETECTION có 4 nhóm tiếp cận sau: Machine Learning, Image Classification, Real-Time, Deep Learning.**

**A. MACHINE LEARNING**

***1. SIFT (Scale-Invariant Feature Transform):*** thuật toán giúp phát hiện đặc trưng (feature detection) trong CV để phát hiện và mô tả các đặc trung local trong hình ảnh, được phát triển bởi David Lowe vào năm 1999. SIFT được nhận định là một kĩ thuật mạnh mẽ để oject recognition/detection. Được áp dụng trong các lĩnh vực object/face recognition, 3D modeling, navigation,… SIFT sẽ trích xuất từ 1 tập hợp các hình ảnh được tham chiếu trong database. Một object được nhận dạng trong một hình ảnh mới bằng cách so khớp từng đặc trưng từ ảnh mới đó với database và tìm ra các ứng cử mà matching được với đặc trưng dựa trên tính khoảng cách Euclidean của các vector đặc trưng.

*- Bài báo: Face recognition using SIFT features*

- Author: C Geng, X Jiang

- Year: 2009

- Cited: 220

-Link: <https://ieeexplore.ieee.org/abstract/document/5413956>

***2. HOG (Histogram of Oriented Gradient):*** HOG là một feature descriptor dùng trong CV và xử lí ảnh nhma82 mục địch object detection. Phương pháp rút trích đặc trưng hình ảnh HOG được xuất bản ở hội nghị CVPR 2005 được đề xuất bởi Dalal và Triggs. HOG lần đầu được đề xuất cho phương pháp rút trích đặc trưng sử dụng các thống kê histogram về hướng trên ảnh gradient cho bài toán human detection. HOG dùng cho object detection với các bước như đọc ảnh; tiền xử lí bằng các thao tác cân bằng sáng, làm mờ; rút trích đặc trưng ảnh (feature extraction): mã hóa hình ảnh thành một vector, vector này mang những đặc trưng cho ảnh đó; huấn luyện mô hình máy học, sử dụng SVM để phân tách các vecor đặc trưng thành các lớp cần phân loại; đánh giá mô hình.

*- Bài báo: Histograms of Oriented Gradients for Human Detection*

- Author: Navneet Dalal, Bill Triggs

- Year: 2005

- Cited: 32964

- Link: <http://lear.inrialpes.fr/people/triggs/pubs/Dalal-cvpr05.pdf>

***3. VIOLA-JONES (Viola-Jones object detection framework):*** là framework của object detection đầu tiên cung cấp tỉ lệ phát hiện đối tượng cạnh tranh trong real-time do Paul Viola và Michael Jones đề xuất năm 2001. Nó có thể được huấn luyện để phát hiện nhiều lớp object khác nhau, tuy nhiên tốt nhất vẫn là face detection. Để có thể phát hiện được khuôn mặt, Viola-Jones yêu cầu toàn bộ khuôn mặt phải hướng về máy ảnh và không được nghiêng sang một bên. Thuộc tính chính của Viola-Jones framework là thời gian huấn luyện chậm nhưng detection nhanh. Hiệu quả của thuật toán này có thể tăng lên đáng kể bằng cách tạo ra hình ảnh tích phân đầu tiên.

*- Bài báo: A comparison of face and facial feature detectors based on the Viola-Jones general object detection framework*

- Author: Modesto Catrillon, Oscar Deniz, Daniel Hernandez, Javier Lorenzo

- Year: 2011

- Cited: 138

- Link: <https://link.springer.com/article/10.1007/s00138-010-0250-7>

**B. IMAGE CLASSIFICATION**

***1. MOBLILENET V3***: Mạng mobilenet do Google sáng tạo, sử dụng mô hình bằng cách tính tích chập có tên Depthwise Separable Convolution để giảm kích thước mô hình và giảm độ phức tạp tính toán. Mạng moblienetV3 được điều chỉnh cho phù hợp với các CPU của di động thông qua sự kết hợp của tìm kiếm kiến ​​trúc mạng phần cứng (NAS) được bổ sung bởi thuật toán NetAdapt. Hai mô hình MobileNet mới để phát hành: MobileNetV3-Large và MobileNetV3-Small được nhắm mục tiêu cho các trường hợp sử dụng tài nguyên cao và thấp. Các mô hình này sau đó được điều chỉnh và áp dụng cho các nhiệm vụ phát hiện đối tượng.

Các thông số vượt trội của MoblieNetV3 đối với tiền nhiệm:

* MobileNetV3-Large chính xác hơn 3,2% về phân loại ImageNet trong khi giảm độ trễ 20% so với MobileNetV2.
* MobileNetV3-Small chính xác hơn 6,6% so với kiểu MobileNetV2 có độ trễ tương đương.
* Phát hiện MobileNetV3-Large nhanh hơn 25% với độ chính xác gần giống như MobileNetV2 trên COCO detection
* MobileNetV3-Large LRASPP nhanh hơn 34% so với MobileNetV2 R-ASPP với độ chính xác tương tự đối với Cityscapes segmentation.

*- Bài báo: Searching for MobileNetV3*

- Author: Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, Hartwig Adam

- Year: 2019

- Cited: 356

- Link: <https://link.springer.com/article/10.1007/s00138-010-0250-7>

**C. REAL-TIME**

***1. CenterNet*** : là một mạng object detection có thiết kế cực kì đơn giản nhưng lại đạt được cân bằng giữa tốc độ và độ chính xác tốt được ra mắt năm 2019. Hướng tiếp cận của CenterNet là đưa bài toán object detection về bài toán tìm điểm đặc trưng (keypoint estimation), từ đó cũng suy ra kích thước và tính toán được bounding box cho bài toán phát hiện vật. Kiến trúc mạng cũng có thể dễ dàng được sửa lại để output ra vị trí 3D, hướng và tư thế cho các bài toán khác.

Nó vượt qua các thuật toán 1 stage phổ biến nhất hiện nay là YOLO v3, RetinaNet trong sự cân bằng giữa tốc độ và độ chính xác. Hơn nữa độ chính xác của CenterNet còn ngang ngửa Faster RCNN - một mạng phát hiện vật 2 stage.

Sau khi qua mạng backbone, ảnh đầu vào sẽ được biến đổi thành một heatmap (bản đồ nhiệt). Mỗi ô trong bản đồ heatmap này thể hiện xác suất trong ô đó chứa tâm của vật. Tiếp đó CenterNet thực hiện lọc các điểm cực đại trên heatmap để xác định tâm của các vật trên ảnh. Từ đó có thể suy ra được kích thước của vật (với bài toán phát hiện vật) và các đặc điểm khác với các bài toán khác.

**Khuyết điểm:**

+ Hoạt động kém với các vật có tâm ở gần nhau

+ Thời gian huấn luyện khá lâu

*- Bài báo: Objects as Points*

- Author: Xingyi Zhou, Dequan Wang, Philipp Krahenbuhl

- Year: 2019

- Cited: 264

- Link: <https://arxiv.org/abs/1904.07850>

*- Bài báo: Centernet: Keypoint triplets for object detection*

- Author: Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, Qi Tian

- Year: 2019

- Cited: 208

- Link:

<https://openaccess.thecvf.com/content_ICCV_2019/html/Duan_CenterNet_Keypoint_Triplets_for_Object_Detection_ICCV_2019_paper.html>

***2. RetinaNet***: sử dụng mạng kim tự tháp tính năng để phát hiện hiệu quả các đối tượng ở nhiều quy mô và giới thiệu một mất mát mới, chức năng mất tiêu điểm, để giảm bớt vấn đề mất cân bằng lớp nền trước-hậu cảnh cực đoan.

Focal loss là hàm loss function lần đầu được giới thiệu trong RetinaNet. Hàm loss function này đã chứng minh được tính hiệu quả trong các bài toán object detection. Đây là lớp bài toán có sự mất cân bằng nghiêm trọng giữa hai class positive (các bounding box có chứa object) và negative (các bounding box chứa object). Thường thì negative có số lượng lớn hơn positive rất nhiều.

**Ưu điểm :** phổ biến, chính xác và chạy nhanh.

*- Bài báo: Focal loss for dense object detection*

- Author: [Tsung-Yi Lin](https://arxiv.org/search/cs?searchtype=author&query=Lin%2C+T), [Priya Goyal](https://arxiv.org/search/cs?searchtype=author&query=Goyal%2C+P), [Ross Girshick](https://arxiv.org/search/cs?searchtype=author&query=Girshick%2C+R), [Kaiming He](https://arxiv.org/search/cs?searchtype=author&query=He%2C+K), [Piotr Dollár](https://arxiv.org/search/cs?searchtype=author&query=Doll%C3%A1r%2C+P)

- Year: 2017

- Cited: 4696

- Link: <https://arxiv.org/abs/1708.02002>

**D. DEEP LEARNING**

***1. One-stage***: Object Detection với 1 số model điển hình như: SSD, Yolo v2/3, RetinaNet. Gọi là one-stage vì trong việc thiết kế model hoàn toàn không có phần trích chọn các vùng đặc trưng (các vùng có khả năng chứa đối tượng) như RPN của Faster-RCNN. Các mô hình one-stage OD coi phần việc phát hiện đối tượng (object localization) như 1 bài toán regression (với 4 tọa độ offset, ví dụ x, y, w, h) và cũng dựa trên các pre-define box gọi là anchor để làm việc đó. Các mô hình dạng này thường nhanh hơn tuy nhiên "độ chính xác" của model thường kém hơn so với two-stage object detection.

**Ưu điểm :** 1 số mô hình one-stage vẫn tỏ ra vượt trội hơn 1 chút so với two-stage như Retina-Net với việc việc thiết kế mạng theo FPN (Feature Pyramid Network) và Focal Loss

*- Bài báo: FCOS: Fully Convolutional One-Stage Object Detection*

- Author: Zhi Tian, Chunhua Shen, Hao Chen, Tong He

- Year: 2019

- Cited: 307

- Link:

<http://openaccess.thecvf.com/content_ICCV_2019/html/Tian_FCOS_Fully_Convolutional_One-Stage_Object_Detection_ICCV_2019_paper.html>

**2. Two-stage**: Object Detection điển hình như RCNN / Fast-RCNN / Faster-RCNN / Mask-RCNN (object detect + instance segmentation), ... Việc gọi là two-stage là do cách model xử lý để lấy ra được các vùng có khả năng chứa vật thể từ bức ảnh. Ví dụ, với Faster-RCNN thì trong stage-1, ảnh sẽ được đưa ra 1 sub-network gọi là RPN (Region Proposal Network) với nhiệm vụ extract các vùng trên ảnh có khả năng chứa đối tượng dựa vào các anchor (phần này mình sẽ đề cập kĩ hơn bên dưới). Sau khi đã thu được các vùng đặc trưng từ RPN, model Faster-RCNN sẽ thực hiện tiếp việc phân loại đối tượng và xác định vị trí nhờ vào việc chia làm 2 nhánh tại phần cuối của mô hình (Object classification & Bounding box regression).

*- Bài báo: Light-head R-CNN: In Defense of Two-Stage Object Detector*

- Author: Zeming Li, Chao Peng, Gang Yu, Xiangyu Zhang, Yangdong Deng, Jian Sun

- Year: 2017

- Cited: 142

- Link : <https://arxiv.org/abs/1711.07264>