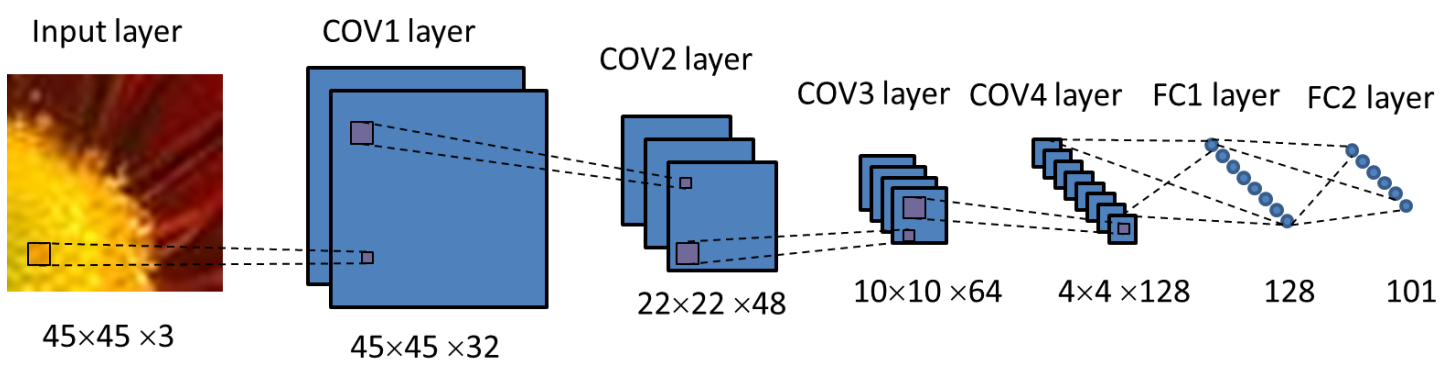
Note reference

1. Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, “You Only Look Once: Unified, Real-Time Object Detection”, Computer Vision and Pattern Recognition (CVPR), IEEE Conferences, pp: 779-788, 2016.
2. Joseph Redmon, Ai Farhadi, “YOLO9000: Better, Faster, Stronger”, Computer Vision and Pattern Recognition (CVPR), IEEE Conferences, pp. 6517-6525, 2017.
3. Lowe, David G, “Distinctive image features from scale-invariant keypoints”, International journal of computer vision, Vol 60, number 2, pp. 91-110, 2004.
4. Dipen Narendra Dalal, Bill Triggs, “Histograms of oriented gradients for human detection”, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), IEEE Conferences, Vol 1, pp. 886-893, 2005.
5. Gedas Bertasius, Jianbo Shi, Lorenzo Torresani, “DeepEdge: A multi-sccale bifurcated deep network for top-down contour detection”, Computer Vision and Pattern Recognition (CVPR), IEEE Conferences, pp. 4380-4389, 2015.
6. Wei Shen, Xinggang Wang, Yan Wang, Xiang Bai, Zhijiang Zhang, “DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection”, Computer Vision and Pattern Recognition (CVPR), IEEE Conferences, pp. 3982-3991, 2015.
7. A. Krizhevsky, I. Sutskever, G. E. Hinton, “Imagenet classification with deep convolutional neural networks”, Advances in neural information processing system, pp. 1097-1105, 2012.
8. Dumitru Erhan, Christian Szegedy, Alexander Toshev, Dragomir Anguelov, “Scalable Object Detection Using Deep Neural Networks”, Computer Vision and Pattern Recognition (CVPR), IEEE Conferences, pp. 2155-2162, 2014.
9. Shipra Ojha, Sachin Sakhare, “Image processing techniques for object tracking in video surveillance – A survey”, International Conference on Pervasive Computing (ICPC), IEEE Conferences, pp. 1-6, 2015.
10. Yan Yang, Xiaodong Wang, Jiande Wu, Haitang Chen, Zhaoyuan Han, “An improved mean shift object tracking algorithm based on ORB feature matching”, The 27th Chinese Control and Decision Conference (CCDC), IEEE Conferences, pp. 4996-4999, 2015.
11. Rosten E, Drummond T, “Fusing points and lines for high performance tracking”, Tenth IEEE International Conference on Computer Vision(ICCV’05), IEEE Conferences, Vol 2, pp.1508-1515, 2005.
12. Michael Calonder, Vincent Lepetit, Christoph Strecha, Pascal Fua, “Brief: binary robust independent elementary features”, European Conference on Computer Vision (ECCV), Springer, pp. 778-792, 2010.
13. Jong-Min Jeong, Tae-Sung Yoon, Jin-Bae Park, “Kalman filter based multiple objects detection-tracking algorithm robust to occlusion”, SICE Annual Conference (SICE), IEEE Conferences, pp. 941-946, 2014.
14. Soonmin Hwang, et al, “Fast multiple objects detection and tracking fusing color camera and 3D LIDAR for intelligent vehicles”, Ubiquitous Robots and Ambient Intelligence (URAI), IEEE Conferences, pp. 234-239, 2016.
15. Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, Philip Torr, “BING: Binarized Normed Gradients for Objectness Esimation at 300fps”, Computer Vision and Pattern Recognition (CVPR), IEEE Conferences, pp. 3286-3293, 2014.
16. C. Lawrence Zitnick, Piotr Dollr, “Edge boxes: Locating object proposals from edges”, European Conference on Computer Vision (ECCV), Springer, pp. 391-405, 2014.
17. Philipp Krhenbhl, Vladlen KoItun, “Geodesic object proposals”, European Conference on Computer Vision (ECCV), Springer, pp. 725-739, 2014.
18. Jasper RR Uijlings, Koen E. A. van de Sande, Theo Gevers, Arnold W. M. Smeulders, “Selective search for object recognition”, International Journal of Computer Vision (IJCV), Springer, Vol 104, Number 2, pp. 154-171, 2013.
19. Ross Girshick, “Fast R-CNN”, IEEE International Conference on Computer Vision (ICCV), IEEE Conferences, pp. 1440-1448, 2015.
20. Latha Anuj, M T Gopala Krishna, “Multiple camera based multiple object tracking under occlusion: A survey”, 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE Conferences, pp. 432-437, 2017.
21. Wei Li, Rui Zhao, Tong Xiao, Xiaogang Wang, “DeepReID: Deep Filter Pairing Neural Network for Person Re-identification”, Computer Vision and Pattern Recognition (CVPR), IEEE Conferences, pp. 152-259, 2014.

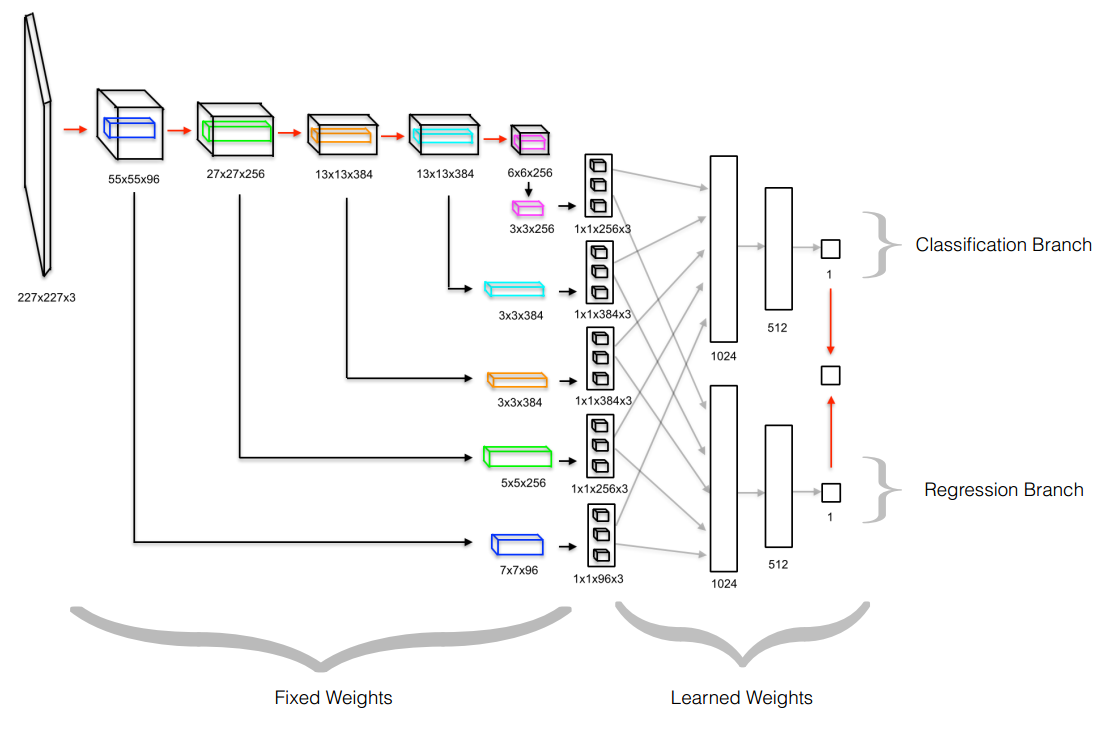
Tác giả Joseph Redmon trong bài nghiên cứu [1] và [2] của mình đã đề xuất một hệ thống phục vụ cho việc phát hiện đối tượng trong frame ảnh có tên YOLO. Tác giả xây dựng hệ thống YOLO là một mạng CNN. Cũng giống các phương pháp phát hiện đối tượng khác như DPM, R-CNN, Fast R-CNN… YOLO có khả năng dự đoán được vị trí của các đối tượng đồng thời phân lớp cho các đối tượng đó dựa trên việc học các đặc trưng của đối tượng như HAAR, SIFT, HOG từ các ảnh trong tập huấn luyện có kích thước đầy đủ. Tác giả chứng minh được rằng YOLO tính toán nhanh với khả năng xử lý 45 frame ảnh trên một giây với bản đầy đủ và 155 frame ảnh trên một giây với bản thu nhỏ. Do đó, YOLO có thể được sử dụng trong phát hiện đối tượng đối với các ứng dụng đòi hỏi tính toán nhanh và đáp ứng thời gian thực. Tuy nhiên, YOLO vẫn chưa thể đáp ứng được độ chính xác cao khi so sánh với các phương pháp phát hiện đối tượng hiện đại nhu Fast R-CNN…

Phát hiện viền được coi là một phương thức cơ bản trong phân mảng, nhận dạng ảnh ảnh và hệ thống phát hiện đối tượng. Phát hiện viền thường sử dụng các đặc trưng như SIFT [3] và HOG [4] của từng pixel trong ảnh để xác định pixel ảnh đang xét có thuộc đường viền hay không. Cách tiếp cận này được sử dụng khá rộng rãi và hộ trợ các giải thuật hang đầu trong bài toán phát hiện viền của đối tượng. Tuy nhiên, không thể phủ nhận rằng với cách tiếp cận trên giải thuật phân lớp không đem lại tính tách biệt cao giữa pixel ảnh thuộc và không thuộc viền. Chính vì vậy mà nhiều nhà khoa học đã sử dụng học đặc trưng sâu (deep features) để giải quyết vấn đề phân lớp giữa pixel thuộc và không thuộc viền. Gedas Bertasius [5] hay Wei Shen [6] xây dựng một mạng nơ ron với sáu lớn, bốn lớp đầu là lớp CL và hai lớp cuối là lớp liên kết đầy đủ FCL. Giá trị đầu vào của mạng CNN mà tác giả Wei Shen đề xuất là một ảnh trong không gian màu RGB với mỗi mảng được chia nhỏ kích thước 45x45 và giá trị đầu ra là một vector 128 chiều được coi như là đặc trưng học sâu sử dụng cho các phương pháp phát hiện viền. Hình 1 mô tả kiến trúc mạng nơ ron mà Wei Shen đề xuất.

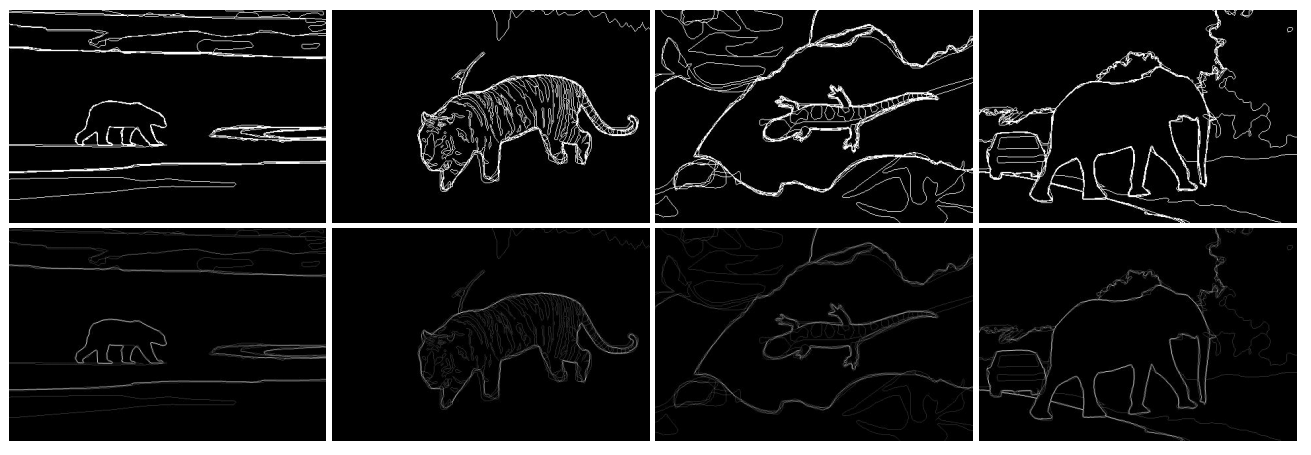


Hình 1: cấu trúc CNN Wei Shen đề xuất [6].

Tác giả Gedas Bertasius [5] lại đề xuất một cách tiếp cận khác cũng dựa vào việc xây dựng CNN. Giá trị đầu vào của cách tiếp cận mà tác giả đề xuất là ảnh cần phát hiện viền. Sử dụng phương pháp phát hiện cạnh của Canny (Canny edge detector) để chọn ra các điểm có khả năng nằm trên cạnh của đối tượng và rút trích ra mảng tương ứng với mỗi điểm được chọn nằm ở trung tâm của mảng đó. Tập các mảng thu được chuyển sang kích thước 227x227x3 để đưa vào mạng KNet [7] rút trích đặc trưng là các mảng có chứa các điểm có khả năng thuộc viền của đối tượng. Các đặc trưng này sau đó được đưa vào mạng con phân nhánh với hai nhánh và mỗi nhanh gồm hai lớp liên kết đầy đủ. Nhánh thứ nhất được huấn luyện để thực hiện phân lớp viền và nhánh thứ hai được huấn luyện để học được sự khác nhau giữa các viền được nhận dạng bởi các phần khác khác nhau. Hình 1 mô tả kiến trúc của mạng CNN mà Gedas đề xuất

Hình 1: Cấu trúc CNN Gedas Bertasius đề xuất [5].

Và hình 2 mô tả giá trị đầu ra của hai nhánh trong mạng CNN trên



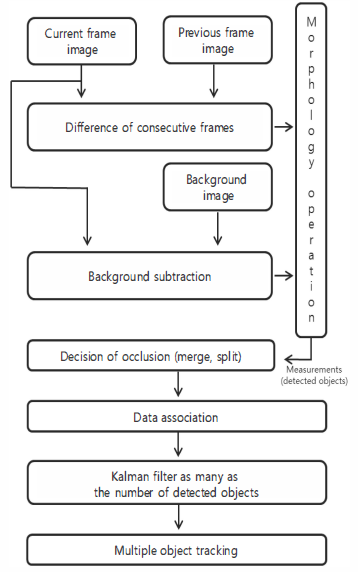
Hình 2: Kết quả đầu ra của nhánh phân lớp (trên) và của nhánh hồi quy (dưới) [5].

Tác giả Dumitru Erhan và các đồng nghiệp [8] đã đề xuất phương pháp phát hiện nhiều đối tượng trong một frame ảnh được gọi là “DeepMultiBox”. Đối với bài toán phát hiện đối tượng trong frame ảnh, ta có rất nhiều cách như là background subtraction và motion detection, supervisor learning… Phương pháp đề xuất này sử dụng DNNs là một trong các giải thuật thuộc lớp supervior learning. Mục tiêu của nghiên cứu là xây dựng được phương pháp dự đoán được tập các vùng chứa đối tượng gọi là bounding box, bounding box là một hình chữ nhật bao quanh đối tượng trong không gian 2D. Dữ liệu đầu ra của phương pháp phát hiện đối tượng được đề xuất bởi Dumitru Erhan và các đồng nghiệp ở đây bao gồm một tập các bounding box với các điểm tọa độ thể hiện vị trí của bounding box trong frame ảnh và giá trị cho biết độ tin cậy (tính chính xác) của việc xác định nhãn của đối tượng tương ứng với bounding box đó. Đóng góp chính của nghiên cứu này là xây dựng được một mạng noron học sâu để phát hiện được đối tượng và thu được dữ liệu đầu ra như mô tả trên.

Tác giả Shipra Ojha trong bài nghiên cứu [9] của mình đã thực hiện một cuộc khảo sát tập trung vào bài toán truy vết đối tượng trong đoạn video quan sát. Với bài nghiên cứu đó, tác giả đã làm rõ nhiều phương thức truy vết thuộc nhiều lớp khác nhau cũng như các chiến lượt nhằm giải quyết bài toán truy vết như dựa vào vùng, viền của đối tương... Đồng thời chỉ ra được điểm tích cực và tiêu cực của các chiến lượt tiếp cận đó. Bài nghiên cứu cũng giới thiệu khá tổng quan về các kiến thức tuy cơ bản nhưng lại hữu ích cho những nghiên cứu về sau tham khảo và đặc biệt là chỉ ra điểm mạnh, điểm yếu của những phương pháp được sử dụng trong truy vết, điều này rất quan trọng cho những nhà nghiên cứu mới tìm hiểu về lĩnh vực thị giác máy tính nói chung và truy vết đối tượng nói riêng.

Tác giả Yan Yang và các đồng nghiệp trong bài nghiên cứu [10] đã đề xuất sử dụng đặc trưng ORB (Oriented FAST and Rotated BRIEF) để cải thiện hiệu suất của phương pháp truy vết đối tượng sử dụng Mean Shift. Giải thuật Mean Shift thông thường sử dụng đặc trưng về màu sắc của đối tượng để truy vết. Các đặc trưng màu ở đây được thu nhận từ không gian màu RGB và chuyển sang không gian màu HSV nhằm giảm bớt sự tác động từ các yếu tố ngoại như ánh sáng…Nhưng với nghiên cứu [10], tác giả sử dụng đặc trưng ORB là một sự cải tiến dựa trên phát hiện đặc trưng FAST [11] và mô tả đặc trưng BRIEF [12]. So với SIFT và SURF thì ORB cải tiến hơn về tốc độ tính toán cũng như đảm bảo tính bất biến của đặc trưng trong các trường hợp các đối tượng bị thay đổi vì xoay, thu phóng hay sự chiếu sáng từ bên ngoài.

Tác giả Jong-Min Jeong trong bài nghiên cứu của mình [13] đã đưa ra giải pháp giải quyết vấn đề phủ lấp giữa các đối tượng di chuyển trong camera. Rõ rang trong thực tế, khi các đối tượng di chuyển qua lại trong camera ngẫu nhiên không theo một hướng nhất định thì việc hai đối tượng che phủ lẫn nhau trong camera là rất thường xuyên. Khi các đối tượng chồng lấp lên nhau như vậy, ta không thể sử dụng các phương pháp phát hiện đối tượng như background subtraction và motion information, supervisor learning để xác định đối tượng bị che phủ đằng sau được mà chỉ có thể sử dụng các phương pháp ước lượng, phỏng đoán. Do đó, trong nghiên cứu [13] tác giả Jong-Min Jeong và các đồng nghiệp đã sử dụng kalman filter và đề xuất phương pháp của mình nhằm giải quyết bài toán che phủ giữa các đối tượng. Đầu tiên, họ sử dụng background subtraction và motion information để phát hiện nhiều đối tượng di chuyển trong camera. Sau đó, họ xác định được số lượng các đối tượng di chuyển trong frame. Bước thứ hai, họ sử dụng Kalman Filter cho mỗi đối tượng ghi nhận được. Tuy nhiên, việc sử dụng một Kalman Filter cho một đối tượng ghi nhận được sẽ dẫn đến tình trạng ở frame ảnh tiếp theo họ không thể biết chính xác được đối tượng nào sẽ tương ứng với bộ Kalman Filter nào trước đó. Chính vì thế, họ đề xuất giải thuật xác định đối tượng ghi nhận và bộ Kalman Filter đúng của nó sử dụng hàm chi phí bao gồm các đặc trưng cũng như là xác định được hai đối tượng che phủ hợp nhất lại với nhau hay tách rời nhau. Hình 1 thể hiện các bước trong phương pháp đề xuất của tác giả và các đồng nghiệp.



Hình 1: Sơ đồ khối phương pháp đề xuất [13].

Hai bước quan trọng mà tác giả và các đồng nghiệp đề xuất để giải quyết được bài toán truy vết các đối tượng bị che phủ lẫn nhau là bước xác định các đối tượng che phủ đang hợp nhất lại với nhau hay đang tách ra và bước gán đối tượng phát hiện được trong frame ảnh tiếp theo đúng với bộ Kalman Filter của nó trong frame ảnh trước. Để phát hiện được các đối tượng đang hợp nhất hay tách rời nhau trong vùng che phủ, họ sử dụng tỉ lệ giữ chiều cao và chiều rộng của đối tượng phát hiện được so sánh với ngưỡng đề xuất. Cụ thể:

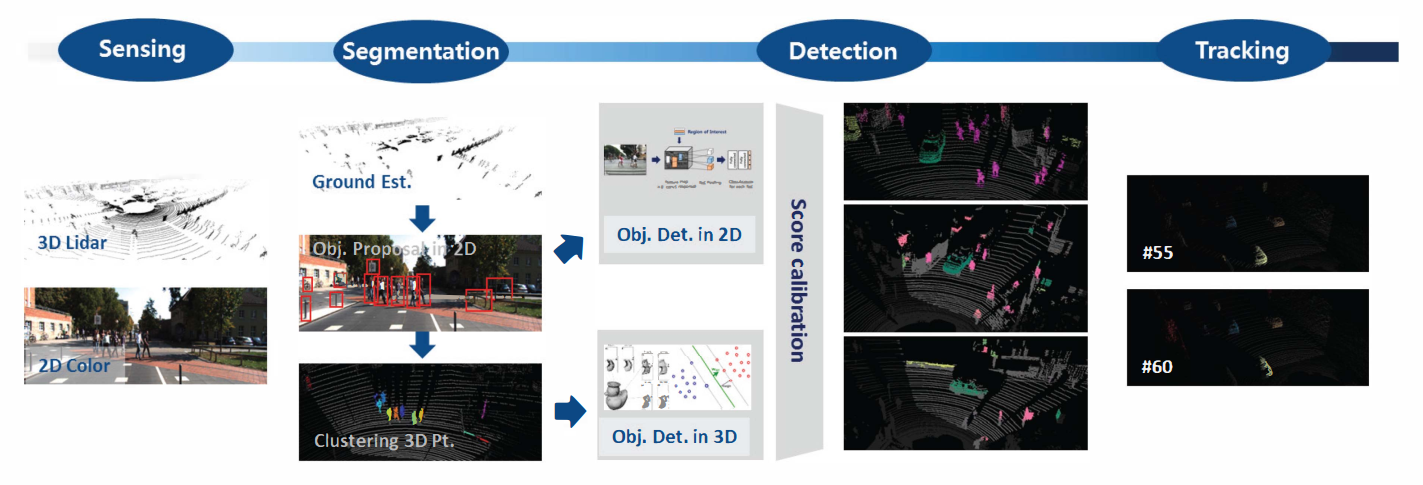
Với m: số lượng các đối tượng phát hiện được trong frame ảnh thứ k.

k: frame ảnh thứ k.

và là ngưỡng trên và ngưỡng dưới của tỉ lệ giữa chiều cao và chiều rộng của đối tượng phát hiện được trong frame ảnh.

Bước quan trọng thứ hai là bước làm thế nào để xác định đúng đối tượng ghi nhận được với bộ Kalman Filter tương ứng của nó. Để thực hiện được điều này, họ sử dụng hai yếu tố là yếu tố về khoảng cách giữa giá trị dự đoán với giá trị của đối tượng ghi nhận được và yếu tố về diện tích của đối tượng giữa các frame ảnh với nhau.

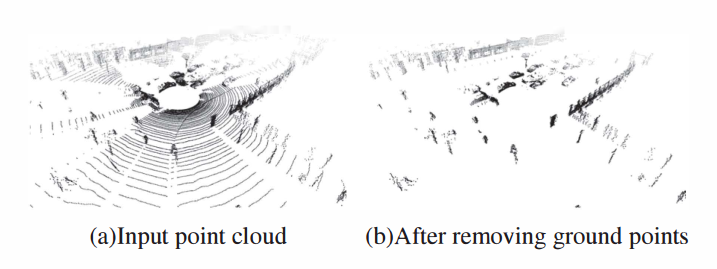
Tác giả Soonmin Hwang [14] đã đề xuất một thư viện cho việc nhận diện và truy vết nhiều đối tượng dựa trên các thông tin từ camera màu và 3D LIDAR (một công nghệ quét laze). Thư viện mà tác giả đề xuất có tốc đồ xử lý nhanh, phản hồi trong thời gian thực nên có thể sử dụng trong các thiết bị thông minh, rô bốt trong lĩnh vực giao thông. Hình 1 giới thiệu các bước trong thư viện mà tác giả đã đề xuất.



Hình 1: Tổng quan về các bước trong phương pháp mà Soonmin Hwang đề xuất [14]

Theo đó, thư viện đề xuất xử lý dữ liệu thông qua bốn bước.

* Bước 1 (Sensing) lấy dữ liệu đầu vào là các ảnh thu được từ camera màu và điểm ảnh 3D thu được từ 3D LIDAR.
* Bước 2 (Segmentation):
  + Phỏng đoán sơ đồ địa hình: từ các điểm ảnh 3D, sử dụng các phương pháp phỏng đoán sơ đồ địa hình để loại bỏ các điểm địa hình như hình 2.



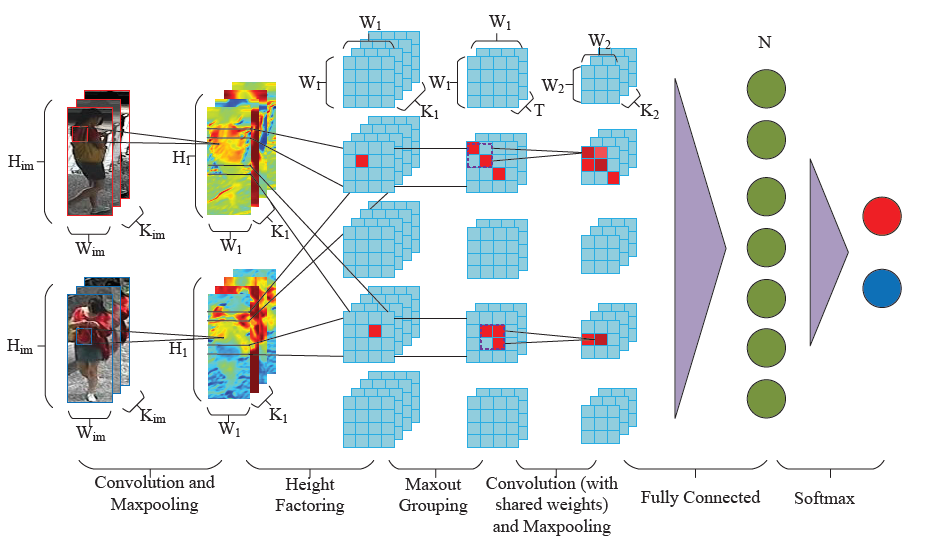
Hình 2: kết quả đầu vào và đầu ra của quá trình phỏng đoán sơ đồ địa hình [14].

* + Đề xuất các đối tượng xuất hiện trong ảnh: với đầu vào là ảnh thu được của camera màu, sử dụng các phương pháp đề xuất đối tượng xuất hiện trong ảnh như BING [15], EdgeBox [16], Geodesic [17], Selective Search [18] để thu được tọa độ và kích thước của các đối tượng có thể có trong ảnh.
  + Gom cụm các điểm ảnh 3D: lấy dữ liệu từ bước bước phỏng đoán sơ đồ địa hình và bước đề xuất đối tượng xuất hiện trong ảnh, bước này sử dụng giải thuật gom cụm DBSCAN để gom các điểm ảnh 3D từ danh sách các điểm ảnh thu được từ giai đoạn phỏng đoán sơ đồ địa hình để loại bỏ các điểm nhiều và chỉ giữ lại các điểm ảnh có khả năng là điểm ảnh thuộc đối tượng. Để tăng độ chính xác của DBSCAN, họ sử dụng thêm kết quả từ giai đoạn đề xuất đối tượng để giữ lại các điểm ảnh có xác xuất cao thuộc đối tượng.
* Bước 3 (detection): Sử dụng Fast R-CNN [19] để xác định đối tượng trong ảnh 2D và sử dụng linear SVN để huấn luyện các điểm ảnh 3D. Từ hai kết quả rời rạc, họ tính giá trị tin cậy cho quá trình phát hiện đối tượng của mình.
* Bước 4 (Tracking): rút trích năm đặc trưng như giá trị trung bình tọa độ điểm ảnh 3D, phương sai, sơ đồ màu, kích thước của mảng và số lượng điểm ảnh 3D chứa trong đối tượng. Sử dụng khoảng cách Euclid để tính được độ tương đồng của các đối tượng giữa hai frame ảnh liên tiếp nhau.

Thư viện mà tác giả xây dựng có thể giải quyết các trường hợp thách thức ví dụ như bài toán che phủ mà một camera hoặc 3D LIDAR không thể giải quyết độc lập được. Đồng thời với độ chính xác cao và khả năng tính toán trong thời gian thực, thư viện rất kì vọng có thể đưa vào ứng dụng trong các phương tiện di chuyển thông minh.

Tác giả Latha Anuj, trong bài nghiên cứu [20] đã thực hiện một cuộc khảo sát tập trung vào việc giải quyết bài toán che phủ các đối tượng trong truy vết. Tác giả đã phân loạt khá đầy đủ các trường hợp có thể xảy ra trong quá trình truy vết đối tượng như là đối tượng không bị che phủ, bị che phủ một phần, bị che phủ hoàn toàn và bị che phủ hoàn toàn trong một khoảng thời gian dài, hay đối tượng rời khỏi vùng quan sát trong hệ thống nhiều camera cũng dược coi là một trạng thái bị che phủ. Theo những nghiên cứu của mình Ms. Latha xét thấy rằng để giải quyết bài toán che phủ thì việc sử dụng camera đơn không mang lại hiệu quả cao bởi lẽ các cách tiếp cận hiện tại trong việc truy vết đối tượng hầu hết đều dựa vào kết quả của quá trình phát hiện đối tượng mà trong trường hợp che phủ ta không thể thu được. Do đó, việc sử dụng hệ thống nhiều camera sẽ mang lại kết quả tốt hơn rõ rệt. Với việc sử dụng nhiều camera cùng tham gia theo dõi, ta có thể thu được các đặc trưng về độ sâu, vị trí, kết cấu và màu sắc của đối tượng, ngay cả trong trường hợp bị che phủ, ta cũng có thể sử dụng đặc trưng về không gian, thời gian và góc nhìn chia sẻ dược giữa các camera.

Tác giả Wei Li và các đồng nghiệp của mình trong nghiên cứu [21] đã đề xuất xây dung một mạng nơ ron học sâu nhằm xác định định danh của người xuất hiện trong đoạn camera. Ở nghiên cứu [21] tác giả đề xuất mạng nơ ron bộ loc bắt cặp (FPNN – Filter Pairing Neural Network) để xác định định danh của đối tượng di chuyển thường được sử dụng trong các hệ thống camera phân tán không có vùng trùng lắp hay bất cứ thông tin liên hệ trực tiếp với nhau dựa trên các đặc trưng về màu sắc của đối tượng. Điểm đóng góp của nghiên cứu này đó là thay vì sử dụng các đặc trưng thủ công bằng thao tác trực tiếp thì tác giả đề xuất phương pháp thu và học các đặc trưng từ dữ liệu có được một cách tự động. mạng nơ ron bộ lặc bắt cặp được mô tả như hình 1:



Hình 1: Filter pairing neural network [21].