**NGHIÊN CỨU THUẬT TOÁN THỊ GIÁC MÁY TÍNH ĐỂ PHÁT HIỆN KHOẢNG CÁCH THEO THỜI GIAN THỰC**

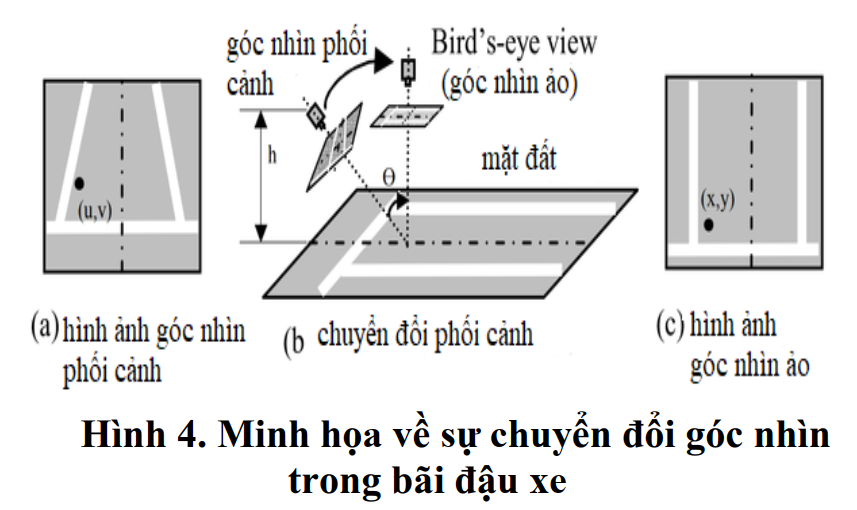
**(Sử dụng Slide anh nhớ lưu ý ở slide số 4 và số 6: Tại slide số 4 click vào Conlutional layers và Full-connected layers nó sẽ zoom tới chi tiết. Tương tự ở slide 6 khi click vào The network view)**

**Ý tưởng:**

Sử dụng lớp mô hình họ YOLO có sẵn trên framework Darknet để phân đoạn và nhận dạng đối tượng. Từ tọa độ có được, sử dụng phép chuyển đổi Bird’s-Eye view để đưa về góc nhìn từ trên xuống, sau đó tính khoảng cách bằng công thức Euclid.

**1. Tính khoảng cách vật thể với phép chuyển đổi Bird’s-Eye view**

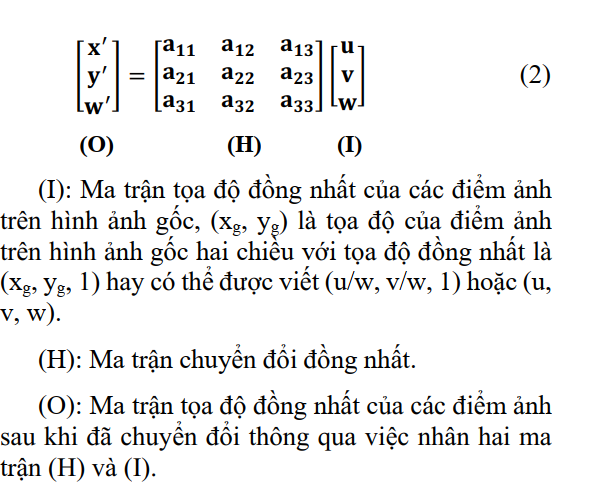
Camera thường quan sát ở góc nghiêng, do đó, tọa độ của các đối tượng quan sát bị ảnh hưởng bởi khoảng cách đến camera, vì vậy nếu tính toán khoảng cách dựa trên tọa độ này thì sai số rất lớn. Phép chuyển đổi Bird’s-Eye view cơ bản là chuyển từ góc nhìn nghiêng về góc nhìn trên xuống, từ đó việc tính toán khoảng cách giữa các đối tượng sẽ ít sai số hơn.



Ý tưởng về quy trình chuyển đổi Bird’s-Eye view như sau:

+ đầu tiên xây dựng ma trận chuyển đổi bằng cách:

Trong hình 4, Mối quan hệ giữa điểm (x, y) của hình ảnh xem qua bird’s-eye view và điểm (u, v) của hình ảnh gốc có thể được xây dựng bằng ma trận đồng nhất (homography matrix) 3 x 3 như sau: (x = ; Y = )

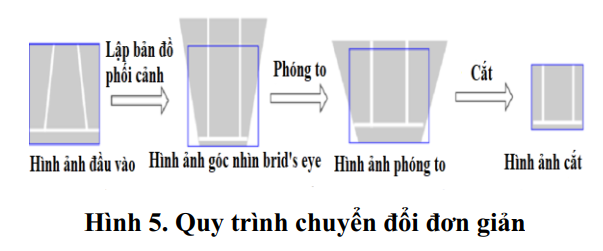


+ Sau khi có ma trận chuyển đổi, ta thực hiện tiếp 3 bước:

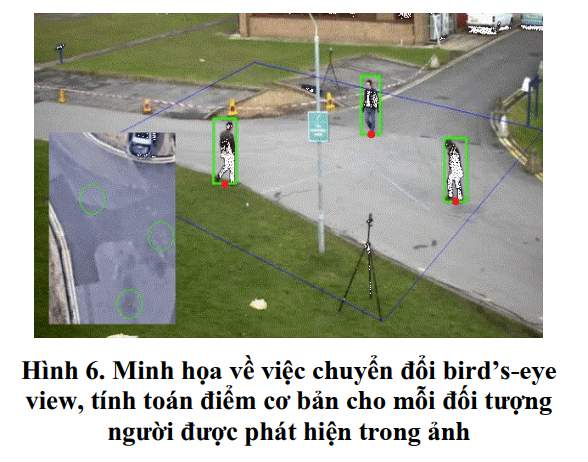
B1: Lập bản đồ phối cảnh

B2: Phóng to

B3: Cắt => thu được hình ảnh theo góc nhìn Bird’s-Eye

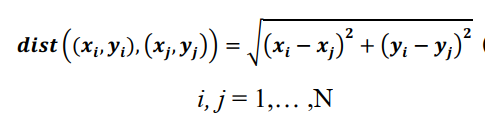


Ví dụ minh họa:



- Xây dựng hộp bao quanh bằng cách lấy 2 điểm: trên cùng bên trái và dưới cùng bên phải. Ta lấy tọa độ của điểm nằm ở tâm dưới cùng của hộp là đại diện cho tọa độ của 1 người.

- Từ tọa độ có được, tính khoảng cách bằng công thức Eculid:



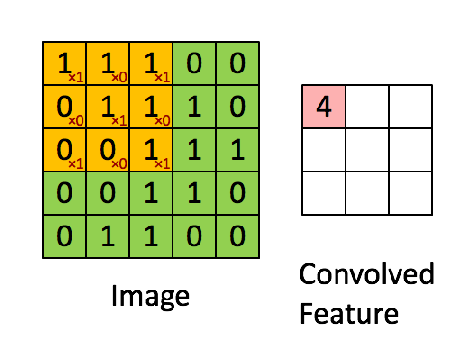
+ Ví dụ: Tính khoảng cách từ điểm A(3, 9) đến B(5,7) ta có:

D =

***2. Trong bài này ta sẽ đi sâu về phương pháp nhận diện đối tượng***

Trong bài này sử dụng nhận diện đối tượng bằng lớp mô hình họ YOLO có sẵn trong framework Darknet.

- Phương pháp: Dựa trên mạng neural network duy nhất được huấn luyện dạng end-to-end model. Mô hình lấy input là 1 bức ảnh và dự đoán các bounding box (hộp bao quanh đối tượng) và nhãn lớp đi kèm. Yolo được tạo ra từ việc kết hợp giữa các convolutional layers và connected layers.Trong đóp các convolutional layers sẽ trích xuất ra các feature của ảnh, còn full-connected layers sẽ dự đoán ra xác suất đó và tọa độ của đối tượng.



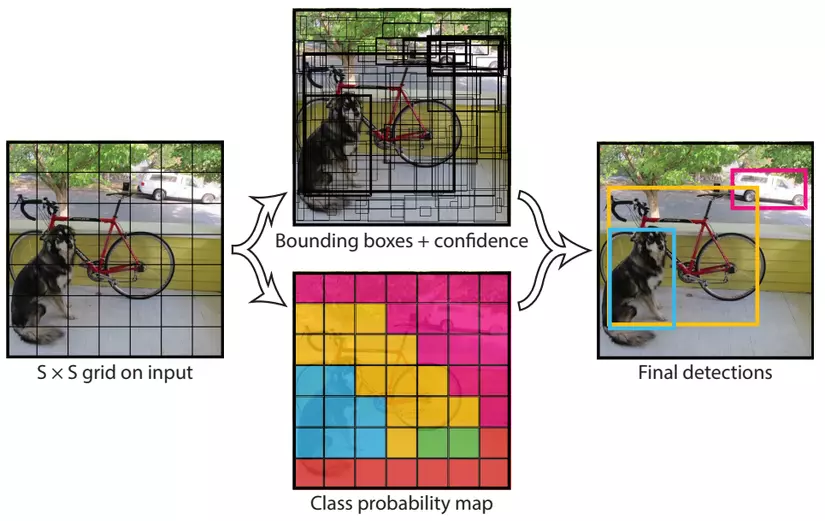
- Mô hình hoạt động:

+ Đầu tiên phân chia hình ảnh đầu vào thành 1 lưới các ô (grid of cell), trong đó mỗi ô chịu trách nhiệm dự đoán các bounding boxes nếu tâm của nó nằm trong ô.

+ Mỗi grid cell (tức 1 ô bất kỳ nằm trong lưới ô) dự đoán các bounding boxes được xác định dựa trên:

* Tọa độ x, y; chiều rộng (w) ; chiều cao (h)
* Độ tin cậy (confidence score) về khả năng chưa vật thể bên trong
* Dự đoán nhãn các đối tượng

- Cụ thể:



Với Input là 1 ảnh, đầu ra mô hình là một ma trận 3 chiều có kích thước S×S×(5×N+M) với số lượng tham số mỗi ô là (5×N+M) với N và M lần lượt là số lượng Box và Class mà mỗi ô cần dự đoán. Ví dụ với hình ảnh trên chia thành

7×7 ô, mỗi ô cần dự đóan 2 bounding box và 3 object : con chó, ô tô, xe đạp thì output là 7×7×13, mỗi ô sẽ có 13 tham số, kết quả trả về (7×7×2=98) bounding box.

Giải thích về con số (5×N+M):

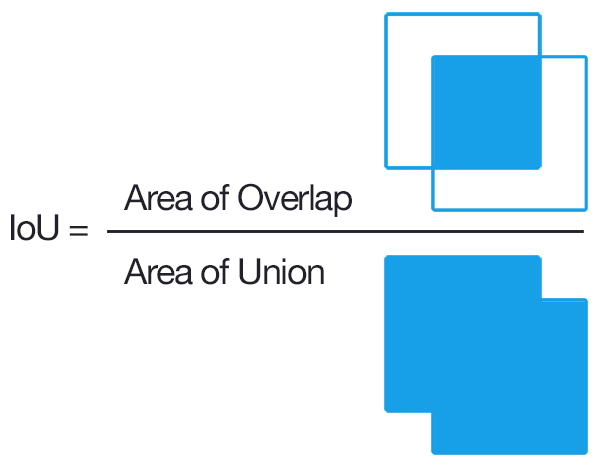
Dự đoán mỗi bounding box gồm 5 thành phần : (x, y, w, h, prediction) với (x, y ) là tọa độ tâm của bounding box, (w, h) lần lượt là chiều rộng và chiều cao của bounding box, prediction được định nghĩa

Với hình ảnh trên như ta tính mỗi ô sẽ có 13 tham số, ta có thể hiểu đơn giản như sau tham số thứ 1 sẽ chỉ ra ô đó có chứa đối tượng nào hay không P(Object), tham số 2, 3, 4, 5 sẽ trả về x, y ,w, h của Box1. Tham số 6, 7, 8, 9, 10 tương tự sẽ Box2, tham số 11, 12, 13 lần lượt là xác suất ô đó có chứa object1( P(chó|object), object2(P(ô tô|object)), object3(P( xe đạp|object)). Lưu ý rằng tâm của bounding box nằm ở ô nào thì ô đó sẽ chứa đối tượng, cho dù đối tượng có thể ở các ô khác thì cũng sẽ trả về là 0. Vì vậy việc mà 1 ô chứa 2 hay nhiều tâm của bouding box hay đối tượng thì sẽ không thể detect được, đó là một hạn chế của mô hình YOLO1, vậy ta cần phải tăng số lượng ô chia trong 1 ảnh lên đó là lí do vì sao mình nói việc chia ô có thể làm ảnh hưởng tới việc mô hình phát hiện đối tượng.



+ Về prediction: như trên có đề cập:



Hàm là hàm đánh giá độ chính xác của object detector trên tập dữ liệu cụ thể. IOU được tính bằng:

Trong đó Area of Overlap là diện tích phần giao nhau giữa predicted bounding box với grouth-truth bouding box , còn Area of Union là diện tích phần hợp giữa predicted bounding box với grouth-truth bounding box. Những bounding box được đánh nhãn bằng tay trong tập traing set và test set. Nếu IOU > 0.5 thì prediction được đánh giá là tốt.