Chương này sẽ trình bày các nghiên cứu có liên quan, khái niệm về thị giác máy tính, mô hình nơ-ron tích chập, lý thuyết về các mô hình phát hiện đối tượng và phương pháp đánh giá mô hình.

## 0.1 Các nghiên cứu liên quan

Yang và Thung [**?** ] đã so sánh mô hình Support Vector Model (SVM) với mô hình mạng nơ-ron tích chập (CNN) dựa trên tập dữ liệu khoảng 2400 hình ảnh và phân ra làm 6 lớp: thủy tinh, giấy, kim loại, nhựa, bìa cứng, và các loại rác khác. Nhóm nghiên cứu đã sử dụng tăng cường dữ liệu bằng việc xoay, điều chỉnh độ sáng, dịch chuyển và cắt các hình ảnh một cách ngẫu nhiên. Sau đó nhóm huấn luyện mô hình và đạt kết quả 67% cho SVM và 22% cho CNN. Tuy kết quả của nhóm tác giả không đạt mong muốn như kì vọng nhưng bộ dữ liệu TrashNet sau này của nhóm đã được sử dụng rộng rãi cho các nghiên cứu tương tự.

Ngoài ra, Jash và Sagar [**?** ] đã đề xuất một mô hình CNN dựa trên 25.077 hình ảnh chia làm 2 lớp: rác hữu cơ và rác tái chế. Mô hình nhận ảnh đầu vào là hình có kích thước 224x224 với mã màu RGB, gồm 6 lớp tích chập, 3 lớp tổng hợp, và 3 lớp kết nối. ReLU sẽ là hàm kích hoạt. Đầu ra của mô hình là nơ-ron có giá trị 0 (rác hữu cơ) và 1 (rác tái chế). Nhóm nghiên cứu so sánh kết quả của mô hình tự xây dựng với VGG16 [**?** ] và ResNet-34. Mô hình đạt độ chính xác lên đến 0.9496.

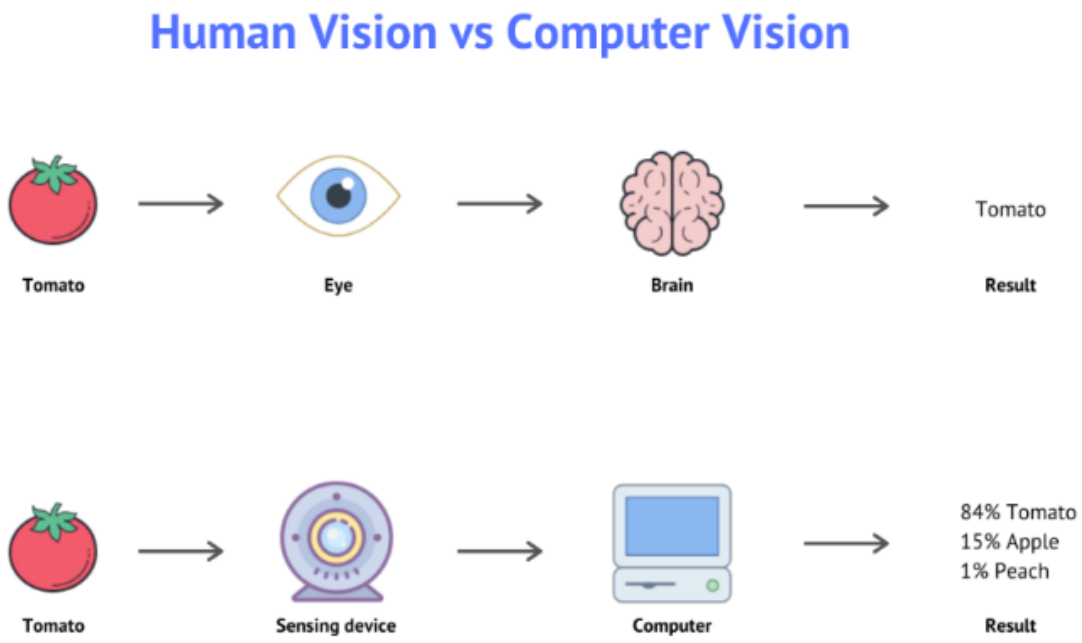
Ahmad và ctv. [**?** ] đã áp dụng các phương pháp kết hợp sớm, kết hợp muộn và kết hợp kép các mô hình học sâu trên tập dự liệu 2.527 hình ảnh đuược cung cấp bởi Kaggle, tạo dữ liệu được chia làm 6 lớp giống như nghiên cứu của Yang và Thung [**?** ]. Trong nghiên cứu này, nhóm tác giả đạt kết quả cao và đưa ra kết luận có thể đạt được hiệu suất phân loại chất thải tốt hơn bằng cách kết hợp tối ưu các mô hình học sâu thông qua các phương pháp thích hợp. Tuy nhiên nhóm nghiên cứu cũng đang mở rộng quy mô dữ liệu trong tương lai.

Ở Việt Nam, nhóm nghiên cứu đến từ Đại học Quốc gia Hà Nội, cục Viễn Thám Quốc gia và viện Khoa học Đo đạc và Bản Đồ thuộc Bộ Tài Nguyên và Môi Trường [**?** ] đã sử dụng mô hình mạng nơ-ron tích chập sâu (DCNN) được huấn luyện và thử nghiệm trên 95 hình ảnh được chụp từ máy bay không người lái (UAV) Phantom 4 Pro ở khu vực ven biển Hội An (Quảng Nam). Mô hình đạt độ chính xác trong việc phân loại hình ảnh rác thải nhựa ven biển là 87%. Các nghiên cứu trên là tiêu biểu của việc ứng dụng công nghệ phát hiện và phân loại rác thải.

## 0.2 Thị giác máy tính

Thị giác máy tính là một trong những lĩnh vực quan trọng của khoa học máy tính và trí tuệ nhân tạo. Thị giác máy tính là một công nghệ giúp tự động nhân biết và mô tả hình ảnh một các chính xác và hiệu quả. Hình 1 giải thích quá trình con người ghi lại hình ảnh đối tượng thông qua võng mạc của mắt, sau đó bộ não tiếp nhận và nhận dạng ra đối tượng. Một người có thể dễ dàng nhận biết và phát hiện đối tượng trong bức ảnh một các chính xác vị trí của chúng. Tuy nhiên việc này lại khó khăn với máy tính, hệ thống phải tiếp nhận hình ảnh thông qua thiết bị ghi hình, "đọc" và "hiểu" hình ảnh dưới dạng ma trận số của tập hợp các điểm ảnh, sau đó được mô hình huấn luyện từ trước để nhận dạng các đối tượng trong ảnh. Tuy vẫn chưa thể chính xác được

như thị giác của con người nhưng đã có rất nhiều ứng dụng hữu ích, điểm hình như điểm danh bằng khuôn mặt, phát hiện các bệnh bằng chuẩn đoán hình ảnh, công nghệ xe tự hành.




Hình 1: Hệ thống thị giác của con người và máy tính

Năm 1966, dự án mang tên "Summer Vision Project" [**?** ] của Seymour Papert và Marvin Minsky đã mở đầu cho việc nghiên cứu về thị giác máy tính sau khi nỗ lực trong hai tháng để tạo ra một hệ thống máy tính có thể nhận dạng các vật thể trong ảnh. Từ đó đến nay, thị giác máy tính đã phát triển vượt bậc để thực hiện được những tác vụ phổ biến như:

- Phân loại hình ảnh: Phân loại hình ảnh cho phép máy tính quan sát và phân loại chính xác một hình ảnh thuộc loại nào. Ví dụ là camera có thể nhận diện khuôn mặt trong ảnh.
- Nhận diện đối tượng: Xác định và phân loại các vật thể khác nhau trong hình ảnh hoặc video bằng ô bao quanh đối tượng (Bbox). Ví dụ phát hiện cây cối hay con người.
- Theo dõi đối tượng: Theo dõi đối tượng sử dụng mô hình học sâu để xác định và theo dõi các mục tiêu. Ví dụ giám sát giao thông tại cái điểm có đèn giao thông.
- Phân đoạn: Xác định đối tượng bằng cách chia nhỏ đối tượng thành các vùng khác nhau dựa trên các điểm ảnh quan sát được. Khác với nhận dạng đối tượng, phân đoạn sẽ xác định hình dạng cụ thể của đối tượng.
- Truy xuất hình ảnh dựa trên nội dung: có khả năng tìm kiếm các hình ảnh kỹ thuật số cụ thể trong cơ sở dữ liệu lớn.

Mạng nơ-ron truyền thống (Neural Network) hoạt động không thực sự hiệu quả với dữ liệu đầu vào là hình ảnh vì các pixel liền kề có sự phụ thuộc lẫn nhau. việc biến đổi thành vector sẽ mất đi tính phụ thuộc, thay đổi ý nghĩa hình ảnh hoặc đòi hỏi dữ liệu lớn để huấn luyện.

## 0.3 Mạng nơ-ron tích chập

Mạng nơ-ron tích chập (CNN) là mạng no-ron phổ biến sử dụng cho dữ liệu ảnh. CNN được thiết kế để tự động học các đặc trưng từ dữ liệu hình ảnh thông qua các tầng tích chập, lần đầu được lấy cảm hứng từ một nghiên cứ năm 1959 của Hubel & Wiesel [**?** ] thông qua phản ứng của các tế bào thần kinh trên não mèo. Với sự kết hợp phát triển đồng bộ và mãnh mẽ của khả năng tính toán của máy tính cùng các phương pháp tối ưu, sau gần 20 năm nghiên cứu, CNN hiện đã và đang phát triển rất nhiều kiến trúc mạng khác nhau. Bắt đầu từ năm 1998 với lần đầu tiên sử dụng mạng tích chập trong tác vụ phân loại chữ số viết tay của Yan Lecun [**?** ] cho đến thành công đầu tiên của mạng AlexNet khi vượt qua được các phương pháp đặc trưng truyền thống như HOG, SHIFT vào năm 2012. Sau đó các mạng mới lần lượt như VGG Net, GoogleNet, ResNet, DenseNet,... đã rút gọn quá trình huấn luyện từ vài ngày xuống còn vài giờ và tăng độ chính xác.

Cấu trúc cơ bản của CNN gồm lớp tích chập, lớp kích hoạt phi tuyến ReLU, lớp gộp và lớp kết nối đầy đủ, được thay đổi về số lượng và cách sắp xếp để tạo ra các mô hình huấn luyện phù hợp cho từng bài toán khác nhau.

### 0.3.1 Tóm tắt các thuật toán phát hiện đối tượng một giai đoạn

Thuật toán phát hiện đối tượng trong một lần chụp hiện là thuật toán phát hiện đối tượng theo thời gian thực nhanh nhất. So với các thuật toán hai giai đoạn, chủ yếu có ba cải tiến. Đầu tiên, thuật toán chụp một lần chuyển nhiệm vụ phát hiện đối tượng thành một bài toán hồi quy, đơn giản hóa quá trình phát hiện đối tượng, cải thiện đáng kể tốc độ phát hiện và đề xuất ý tưởng phát hiện đầu cuối. Thứ hai, thuật toán này đề xuất một phương pháp huấn luyện đa quy mô linh hoạt hơn các thuật toán hai giai đoạn và có thể thích ứng với các bộ dữ liệu khác nhau. Thứ ba, thuật toán chụp một lần cải thiện thiết kế neo của các thuật toán hai giai đoạn và tìm ra cài đặt neo tối ưu bằng cách sử dụng phân cụm.

**Thuật toán 1** Thuật toán NMS

**Đầu vào:** $B = b_1, ..., b_n, P = p_1, ..., p_n, \lambda_p, \lambda_{IoU}$
$B$ là danh sách các hộp giới hạn
$P$ chứa điểm đối tượng tương ứng
$\lambda_p$ xác định ngưỡng đối tượng
$\lambda_{IoU}$ là ngưỡng IoU
**Đầu ra:** $B_r = \{\}, P_r = \{\}$
$B_r$ là danh sách các hộp không triệt tiêu tối đa
$P_r$ chứa điểm đối tượng tương ứng

```
begin
    B_r ← {}
    P_r ← {}
    for b_i in B do
        if b_i < λ_p then
            B ← B − b_i
            P ← P − p_i
        end
    end
    while B ≠ ∅ do
        P_max ← max(P)
        B_max ← max(b_{P_max})
        B_r ← B_r + B_max
        P_r ← P_r + P_max
        B ← B − B_max
        P ← P − P_max
        for b_i in B do
            if IoU(B_max, b_i) > λ_IoU then
                B ← B − b_i
                P ← P − p_i
            end
        end
    end
    return B_r, P_r
end
```