BÁO CÁO ĐỒ ÁN CUỐI KỲ

Lớp : CS114.K21.KHTN

Môn học : Máy học

Giảng viên : PGS.TS.LÊ ĐÌNH DUY, THS. PHẠM NGUYỄN TRƯỜNG AN

I. Mô tả bài toán

1. Đặt vấn đề

   - Mô hình nhận dạng thực vật giúp phân biệt được các loại thực vật phổ biến ở đây em tập trung vào các loại thực vật bóng mát có lượng bao phủ cao, trồng rộng rãi trong làng đại học . Hệ thống cung cấp thông tin các loài thực vật có thể hữu ích cho các nhà thực vật học, nhà công nghiệp, kỹ sư thực phẩm và y bác sĩ. Mô hình sẽ có khả năng nhận dạng thực vật bằng cách sử dụng hình ảnh lá cây của chính nó.

2. Tổng quan bài toán

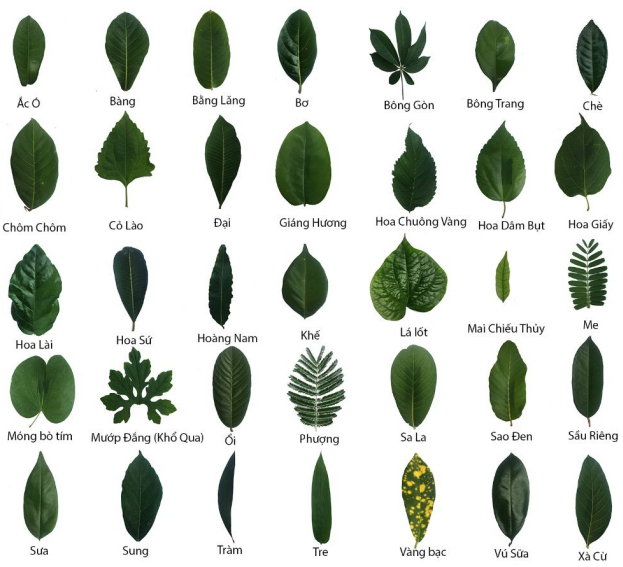
Input : Là một bức ảnh lá cây được chụp từ camera.

Xà Cừ

Output : Đưa ra tên loại lá cây.

II. Xây dựng bộ dữ liệu.

1. Thu thập bộ dữ liệu
    - Tự xây dựng bộ dữ liệu.
    - Thu thập các loại lá cây trong khuôn viên làng đại học , sau đó chụp ảnh bằng camera điện thoại có độ phân giải 3000x4000.
    - Lá cây được chụp chính diện dưới ánh sáng ban ngày , nền chụp trắng.
    - Số lượng gồm 35 loại lá cây với mỗi loại là 20 mẫu lá => Tổng cộng 700 lá.
2. Tiền xử lý dữ liệu
    - Bộ dataset bao gồm 700 ảnh.



- Tất cả hình ảnh các loại lá đều được xoay theo cùng một hướng là đầu lá phải hướng lên trên.
- Em chia làm 2 bộ dataset
    + ORIGIN UIT LEAF DATASET : Bộ dữ liệu chụp chưa cắt background.

+ UIT LEAF DATASET : Bộ dataset đã cắt background trắng.

- Chia thủ công 2 tập train test tương ứng cho mỗi bộ dataset.
- Tỉ lệ train 75% (525 ảnh), test 25% (175 ảnh)
- Bộ test chứa nhiều hình ảnh lá bị biến dạng, như lá non , bị sứt mẻ , lá khô héo, ..vv Nhằm tăng tính phức tạp cho bài toán nhận dạng.

III. Rút trích đặc trưng

- Sử dụng thử nghiệm qua 4 phương pháp

+ Raw feature
+ HOG
+ Perimeter
+ Histogram feature

1. Raw feature
    - Một bức ảnh chiếc lá được đưa về ảnh xám, sau đó scale về bức ảnh có kích thước là 300*400. Đưa ma trận 300*400*1 về một vector có số chiều là 300*400*1. Vector này được sử dụng làm feature cho bức ảnh,hay giá trị của mỗi pixel ảnh được coi là 1 feature.
    - Nhược điểm của phương pháp là làm mất thông tin về không gian giữa các điểm ảnh, dễ bị overfit,không có tính tổng quát , kích thước feature quá lớn gây ra việc tốn kém bộ nhớ . \tài nguyên tính toán và bộ nhớ

+ Khi dùng raw feature chỉ đúng trong trường hợp dữ liệu khớp ở mức độ pixel , gây ra việc chỉ cần sai lệch pixel (ví dụ như xê dịch ảnh) sẽ làm ảnh hưởng tới kết quả.

- Ưu điểm : dễ hiểu , dễ cài đặt. \đơn giản

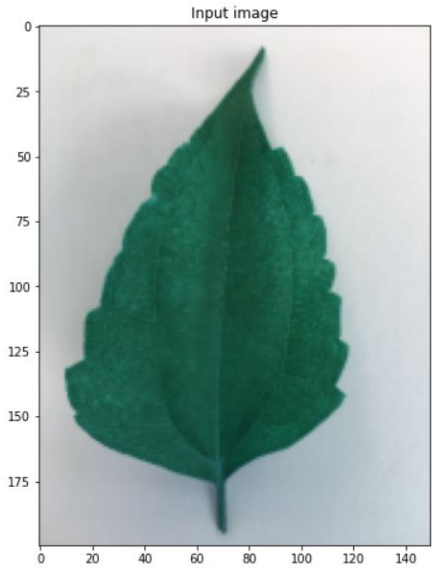
2. HOG

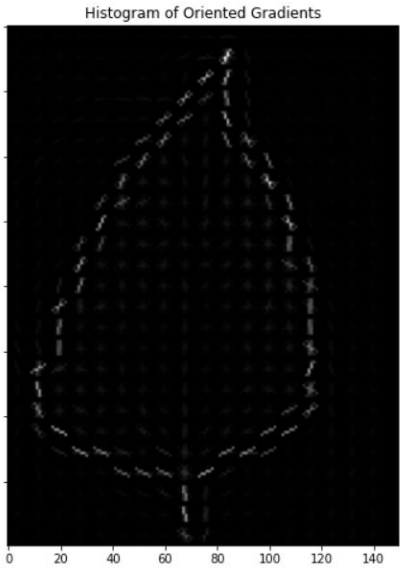
a. Khái niệm và phương pháp sử dụng

- HOG (Histogram of Oriented Gradients) là một loại “feature descriptor” phương pháp sử dụng thông tin về sự phân bố của các cường độ gradient hoặc edge directions để mô tả các đối tượng cục bộ trong ảnh.

  Trong bài toán của em từ một bức ảnh chiếc lá, ta sẽ lấy ra 2 ma trận quan trọng giúp lưu thông tin ảnh đó là độ lớn gradient (gradient magnitude) và phương của gradient (gradient orientation). Một bức ảnh ta sẽ chia thành các block , mỗi block chứa 4 cell có kích thước 8x8 pixel. Sau đó, một biểu đồ histogram thống kê độ lớn gradient được tính toán trên mỗi cell. Tạo thành vector feature của 1 block bằng cách nối liền 4 histogram được tính trên mỗi cell.

  Ví dụ sử dụng HOG trên ảnh lá .



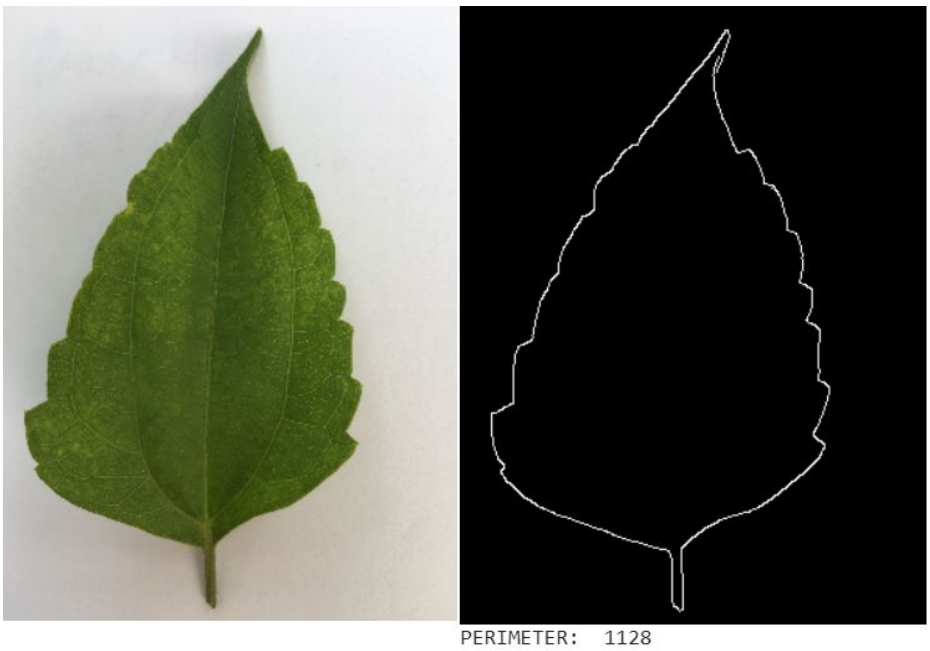


b. Ưu ,nhược điểm của HOG

- Ưu điểm : Vì nó hoạt động trên cell , nó bất biến đối với các phép biến đổi hình học, thay đổi độ sáng. Sử dụng HOG mang lại độ chính xác cao khi sử dụng để tạo đặc trưng đầu vào cho các thuật toán học có giám sát cổ điển, quá trình huấn luyện nhanh và yêu cầu ít tài nguyên tính toán .Đây chính là những ưu điểm vượt trội của HOG so với phương pháp raw feature ở trên.
- Nhược điểm : Việc mạng CNN ra đời đã khiến cho HOG không còn trở nên thông dụng nữa, do khó có thể sử dụng HOG trong bộ dữ liệu lớn.

3. Chu vi

- Sử dụng canny() trong OpenCv để lấy đường viền của mỗi lá . Tính tổng tất cả các điểm cs trên đường viền để chọn ra chu vi là đặc trưng riêng của mỗi lá . Với mỗi lá có chu vi gần giống nhau thì được xem là cùng loại
- Vấn đề xảy ra : Rất thấp so với 2 loại rút đặc trưng còn lại .Vì khi lấy chu vi sẽ mang cách tổng quát , làm mất đi 1 số đặc trưng riêng của từng loại lá ,không phân biệt được các loại lá có chu vi gần bằng nhau.
- Kết luận : Chỉ nên sử dụng feature Perimeter như cách thêm feature phụ để tăng tính ràng buộc của bài toán.

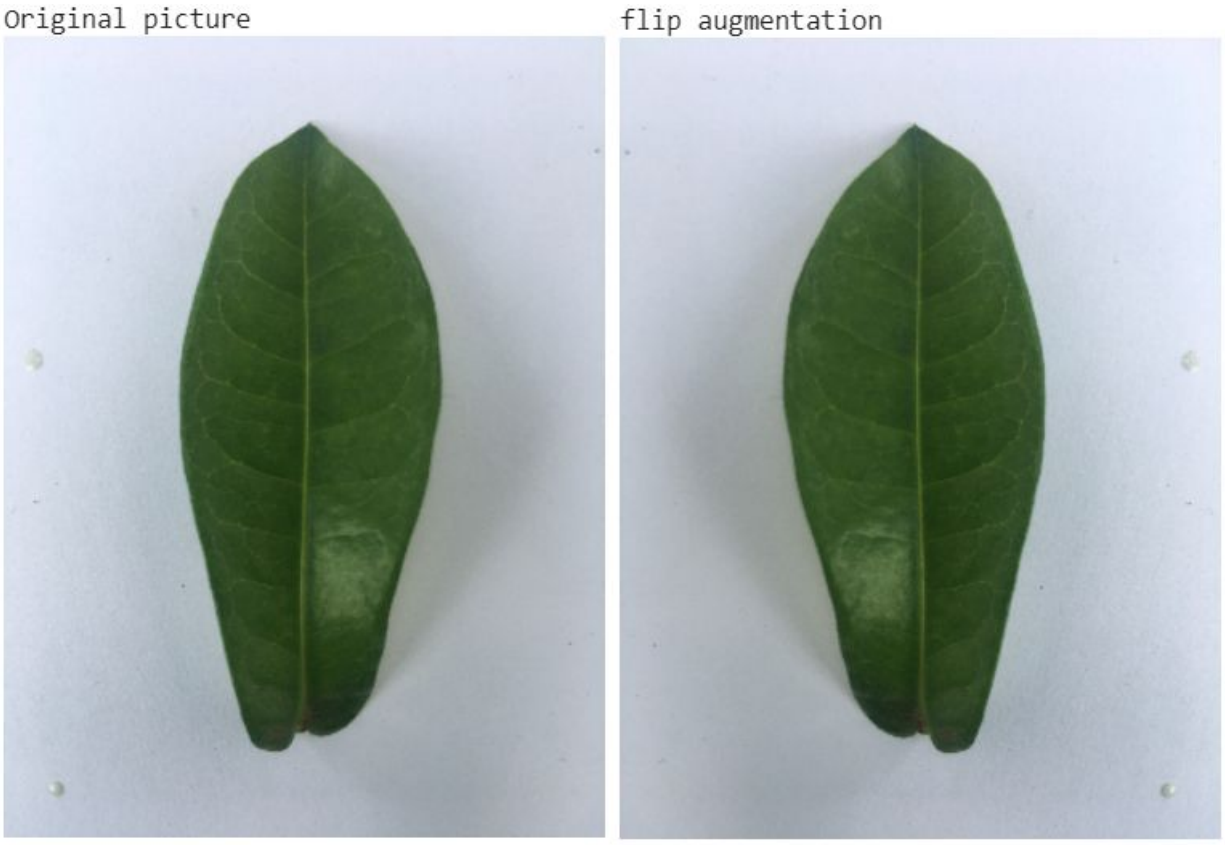


4. Histogram Feature

- Kỹ thuật sử dụng việc phân bổ bảng màu của từng hình ảnh làm feature.
- Không sử dụng được tăng cường dữ liệu .
- Khắc phục làm đặc trưng phụ cho ràng buộc bài toán. Em có kết hợp Histogram Feature và HOG làm tạo thành 1 feature mới cho bài toán . Tuy nhiên việc sử dụng trên làm bài toán tăng hiệu suất lên hơn không đáng kể .Vì Histogram feature chứa lượng vector quá lớn so với HOG , nên sự ảnh hưởng của HOG là rất nhỏ .Dẫn tới việc hiệu quả tăng nhưng không đáng kể.

IV. Tăng cường dữ liệu

- Sử dụng tăng cường dữ liệu trên bộ train data bằng phương pháp flip augmentation.



V. Train model

- Add các model KNN,SVC ,Logistic regression,Random forest Classifier từ thư viện sklearn.
- Thử nghiệm trên 2 bộ dữ liệu ORIGIN UIT LEAF DATASET và UIT LEAF DATASET với 3 phương pháp rút trích đặc trưng kể trên đưa vào 4 model . Sử dụng độ đo F1 và accuracy để đưa ra kết quả cho từng model.

VI. So sánh và đánh giá

- Về bộ dữ liệu , sau thực nghiệm cho thấy bộ dữ liệu ORIGIN UIT LEAF DATASET(bộ dataset gốc) cho hiệu suất cao hơn bộ dữ liệu UIT LEAF DATASET(bộ dataset đã xoá background) với chênh lệch từ 7% ~ 13%.
- Về phương pháp rút trích đặc trưng , em có sử dụng 4 phương pháp là HOG, raw feature , tính chu vi (Shape feature), . Việc sử dụng HOG cho kết quả model huấn luyện đạt hiệu suất cao nhất . Với kết quả thu được sau
-
    + Lấy chu vi lá làm đặc trưng cho hiệu suất 17% ~ 30% với dữ liệu gốc chưa xoá Background .
    + Histogram feature cho hiệu suất từ 57.14% ~ 67% với dữ liệu gốc chưa xoá Background .
    + Với Raw feature và HOG

	Model	Data Original	Data Final
RAW FEATURE	KNN	69	57
	SVC	78	70.29
	LR	76	66.28
	RFC	66	59.43
HOG	KNN	74	61
	SVC	89	81,14
	RL	85	79,42
	RFC	48	58,85

- Về sử dụng mô hình thuật toán máy học : Model đạt hiệu suất cao nhất là SVC với 89% cho đặc trưng HOG và bộ dữ liệu gốc.

VII. Khó khăn và hướng giải quyết.

1. Demo

    - Khi lấy capture ảnh từ camera, việc căn góc của lá sao cho hình ảnh chiếc lá có chiều thẳng đứng , từ dưới lên , có ảnh sáng tương đối tốt , lá đặt ngay ngắn , sau nhiều lần điều chỉnh kết quả mới cho ra đúng.

```
/content/drive/My Drive/CAPTURE_PICTURE
Saved to Captured_photo.jpg
Hoa Dâm Bụt
```

Đây là hình ảnh hoa dâm bụt được đưa vào camera , mô hình đã cho ra kết quả đúng sau nhiều lần chỉnh góc của chiếc lá.

2. Khó khăn

    - Trong bộ dataset có nhiều lá có hình thù tương đồng nhau , khó phân biệt.

- Model chưa nhận dạng được các lá có nhiễu nặng, có hình dạng lá xoay ngược không đúng chiều , ánh sáng không đủ , rách, héo ,...

3. Giải pháp
    - Sử dụng đặc trưng phù hợp với chiếc lá hơn
    - Phát triển model theo hướng sử dụng CNN để rút trích đặc trưng. Dùng deep model để cải thiện phần feature extraction.

VII. Reference

1) Flavia Plant Leaf Recognition System, (http://flavia.sourceforge.net/).
2) M.G. Larese, R. Namias, R.M. Craviotto, M.R. Arango, C. Gallo, P.M. Granitto .Automatic classification of legumes using leaf vein image features.
3) C.H. Arun, W.R. Sam Emmanuel, D.C. Durairaj .Texture feature extraction for identification of medicinal plants and comparison of different classifiers
4) Plant leaf recognition using shape features and color histogram with K-nearest Neighbour Classifier.
TrishenMunisamiaMahessRamsurnaSomveerKishnahaSameerchandPudaruthb