

BÀI THU HOẠCH

CHỦ ĐỀ: TEXT RECOGNITION

A. Thông tin cá nhân:

Huỳnh Thị Mỹ Thanh

MSSV: 19127273

B. Bài thu hoạch:

I. Problem statement:

Text recognition gồm những bài toán nhỏ như:

- Xác định vị trí của chữ (text detection)
- Nhận diện văn bản viết gì (text recognition)
- Text enhancement
- Input: hình ảnh hoặc video là một dãy những hình ảnh với dữ liệu là ma trận, pixel màu (font, ngôn ngữ)
- Output: thông tin văn bản chứa trong ảnh/video input (vị trí của văn bản, văn bản được nhận dạng là gì)

II. Challenge:

Bài toán nhận diện từ ảnh:

- Vật thể phức tạp
- Nguồn sáng phức tạp
- Chất lượng ảnh mờ và nhiễu
- Góc chụp

Sự đa dạng về dữ liệu văn bản:

- Đa dạng font
- Chữ viết tay
- Đa dạng ngôn ngữ
- Sự thay đổi về tỉ lệ chữ

III. Methods

- ❖ Text localization: định vị các thành phần văn bản

Phương pháp:

- Connected component analysis (CCA): có thể được coi là một thuật toán đồ thị, trong đó tập hợp con của các thành phần được kết nối và

dán nhãn giống nhau dựa trên sự giống nhau của tính chất như màu sắc, bố cục, không gian.

- Sliding window classification: các cửa sổ hình ảnh đa tỉ lệ lần lượt duyệt qua các vùng trong ảnh để gom nhóm thành các vùng có chứa văn bản. Sử dụng các phép toán hình thái học, CRF hoặc phương pháp đồ thị.

❖ Text verification: xác minh lại kết quả của bước text detection.

Phương pháp:

- Knowledge – based methods: phương pháp này sử dụng các kiến thức về màu sắc, kích thước và không gian, ... để đặt ra thresholds để xác minh những vùng là văn bản.
- Feature discrimination methods: các đặc trưng khác nhau của văn bản bao gồm cấu trúc, cường độ và các đặc trưng về hình dạng, Wavelet, LBP và mô tả HOG, Gabor strokes và hybrids được sử dụng để thực hiện xác minh vùng văn bản.

❖ Text segmentation: phân đoạn thành các từ, các dòng, các ký tự trước khi đến bước nhận dạng.

Phương pháp:

- Text binarization: nhị phân hóa văn bản để trích xuất các pixel văn bản và loại bỏ các pixel nền. Thuật toán liên quan đến adaptive thresholding, mô hình xác suất và phân cụm.
- Text line segmentation: chuyển đổi một vùng nhiều dòng văn bản thành nhiều vùng con với các dòng văn bản đơn lẻ.
- Character segmentation: tách một vùng văn bản thành nhiều ký tự đơn.

❖ Text recognition: nhận dạng hình ảnh từ các vùng ứng viên thành chuỗi văn bản.

Cấp độ:

- Character recognition: Để nhận dạng các ký tự của cùng một phong chữ duy nhất, các đặc điểm cục bộ thì chỉ cần các bộ phân loại đơn giản thường là LDA. Với trường hợp nhiều phong chữ, hay ký tự bị biến dạng, sự đa dạng bên trong gây khó khăn cho việc mô hình hóa các ký tự của cùng một lớp, một giải pháp là có một bộ phân loại cụ thể cho từng loại.
- Word recognition: cấp độ này tập trung vào nhận diện từ trong văn bản mang ý nghĩa cụ thể hoặc so sánh trong tập dữ liệu học.

- End to end recognition: cấp độ này đưa ra hình ảnh có nền phức tạp làm đầu vào, một hệ thống nhận dạng end to end sẽ có các chức năng localization, detection và recognition để chuyển đổi tất cả các vùng văn bản trong hình ảnh thành chuỗi văn bản.

IV. Conclusion

Việc sử dụng các kỹ thuật thị giác máy tính, học máy và mạng nơ-ron sâu đã giúp cải thiện đáng kể hiệu suất của hệ thống nhận dạng chữ viết. Các phương pháp nhận dạng hiện đại cho phép xử lý tài liệu văn bản hàng loạt nhanh chóng và hiệu quả, từ việc chuyển đổi các tài liệu giấy sang định dạng kỹ thuật số cho đến việc trích xuất thông tin từ các ảnh kỹ thuật số có sẵn. Mặc dù các giải pháp nhận dạng chữ viết hiện đại đã đạt được sự tiến bộ đáng kể, nhưng vẫn còn tồn tại một số thách thức và vấn đề phải giải quyết, như việc nhận dạng chữ viết chất lượng thấp, chữ viết có thể được viết tay hoặc được in bằng nhiều kiểu chữ khác nhau. Tuy nhiên, với sự phát triển của công nghệ, chúng ta có thể kỳ vọng rằng các phương pháp nhận dạng chữ viết sẽ ngày càng được cải tiến và phổ biến hơn trong tương lai.