Signboard text detection and recognition in streaming video

Nguyễn Đình Quân - 20521184, Nguyễn Hùng Phát - 22521074

December 14, 2025

LỜI CẢM ƠN

Trước tiên, em xin gửi lời cảm ơn chân thành và sâu sắc đến Ban Giám hiệu nhà trường và Khoa Khoa học Máy tính đã tạo điều kiện học tập và nghiên cứu thuận lợi trong suốt thời gian em theo học tại Trường Đại học Công nghệ Thông tin.

Em xin bày tỏ lòng biết ơn đặc biệt đến Thầy Đỗ Văn Tiến, đã trực tiếp giảng dạy và tận tình hướng dẫn em trong quá trình thực hiện đề tài khóa luận. Những định hướng, chỉ dẫn rõ ràng cùng sự hỗ trợ quý báu từ thầy đã là tiền đề quan trọng giúp em hoàn thành tốt công việc nghiên cứu và viết báo cáo đúng tiến độ. Em cũng xin cảm ơn thầy vì đã cung cấp tài liệu, giải đáp thắc mắc và luôn tạo môi trường học tập tích cực, hiệu quả.

Trong suốt quá trình thực hiện đề tài, em đã có cơ hội vận dụng những kiến thức nền tảng đã được học, đồng thời tích cực học hỏi, tìm tòi thêm các kiến thức mới. Đây là một trải nghiệm quý báu giúp em trưởng thành hơn trong tư duy và kỹ năng làm việc nghiên cứu.

Mặc dù đã nỗ lực hoàn thành đề tài với tinh thần nghiêm túc và cầu thị, nhưng do hạn chế về thời gian và kinh nghiệm, khóa luận không thể tránh khỏi những thiếu sót. Em rất mong nhận được sự thông cảm, góp ý chân thành từ các thầy cô để em có thể tiếp tục hoàn thiện và phát triển trong tương lai.

Em xin chân thành cảm ơn!

TÓM TẮT KHÓA LUẬN

aaaaa.....

Contents

List of Figures

List of Tables

Chapter 1

TỔNG QUAN

1.1 Đặt vấn đề

Phát hiện và nhận dạng văn bản trong ảnh đời thường (*Scene Text Detection and Recognition* – STDR) là một bài toán quan trọng trong thị giác máy tính, thu hút nhiều sự quan tâm nhờ các ứng dụng rộng rãi như dịch tự động, hỗ trợ dẫn đường, số hóa tài liệu ngoài trời, hay phân tích biển báo giao thông. Với đầu vào là ảnh tĩnh hoặc khung hình video, STDR hướng tới việc xác định vị trí xuất hiện của văn bản và trích xuất chính xác nội dung văn bản đó (Hình 1).

Trong số các dạng văn bản đời thường, **văn bản trên biển hiệu** (*Signboard Text*) có ý nghĩa đặc biệt vì thường chứa thông tin về *tên địa điểm, cơ sở kinh doanh, dịch vụ* hoặc *định danh không gian* trong môi trường đô thị. Do vậy, **phát hiện và nhận dạng văn bản biển hiệu** (*Signboard Text Detection and Recognition*) trở thành một nhánh quan trọng của STDR, có nhiều tiềm năng ứng dụng trong hệ thống dẫn đường thông minh, phân tích thông tin đô thị và xây dựng bản đồ số.

Tuy nhiên, STDR nói chung gặp nhiều thách thức do sự đa dạng của phông chữ, kích thước, hướng và bố cục; văn bản có thể bị nghiêng, cong, chồng chéo hoặc hòa lẫn trong nền phức tạp, cùng các phong cách thiết kế nghệ thuật và đa ngôn ngữ (Hình 2). Đối với tiếng Việt, khó khăn còn lớn hơn do hệ thống dấu (◌́, ◌̀, ◌̂, ◌̃, dấu hỏi, dấu nặng) và các nguyên âm biến thể (ô, ê, â, ă, ơ, ư), làm tăng số lượng ký tự cần nhận dạng và dễ gây nhầm lẫn giữa các chữ có hình dạng gần giống (ví dụ *a* với *â*, *ă*, *á*).

Bên cạnh đó, biển hiệu cũng đa dạng về hình dạng, kích thước, vật liệu và thường xuất hiện ở các vị trí phức tạp trong ảnh (Hình 3), như bị che khuất một phần, bị phản xạ ánh sáng, hoặc nằm trong các bối cảnh đông đúc. Theo hiểu biết của chúng tôi, đến nay mới chỉ có một nghiên cứu [2] tập trung vào phát hiện biển hiệu trên đường phố Việt Nam, trong khi hướng kết hợp *cả phát hiện đối tượng biển hiệu* và *nhận dạng nội dung văn bản trên biển hiệu* vẫn còn ít được khai thác.

Ngoài ra, khi mở rộng từ ảnh tĩnh sang **video hành trình** (dashcam), bài toán còn đối mặt với các thách thức đặc thù như mờ do chuyển động, độ phân giải hạn chế của camera hành trình, cùng sự thay đổi liên tục về ánh sáng và góc nhìn. Những yếu tố này khiến việc phát hiện và nhận dạng văn bản trong video trở nên khó khăn hơn so với ảnh đơn lẻ.

Trong phạm vi khóa luận này, chúng tôi tiến hành khảo sát và tổng hợp các hướng tiếp cận liên quan đến STDR và các phương pháp hiện đại [3,4,5,6]. Trên cơ sở đó, chúng tôi lựa chọn, cài đặt, thực nghiệm và đánh giá một số phương pháp tiên tiến [7,8,9,10] trên tập dữ liệu SignboardText [1]. Đồng thời, tập dữ liệu này được **mở rộng** bằng cách **bổ sung nhãn đối tượng biển hiệu** (thay vì chỉ nhãn văn bản). Tiếp theo, chúng tôi áp dụng các phương pháp vào dữ liệu video camera hành trình trên đường phố Việt Nam, với đầu vào là khung hình chứa biển hiệu và đầu ra là *vị trí biển hiệu cùng nội dung văn bản* tương ứng (Hình 4). Từ văn bản trích xuất được, hệ thống hướng tới việc phát triển ứng dụng minh họa, chẳng hạn như xác định loại hình cơ sở (cửa hàng, nhà hàng, trường học, bệnh viện...), hỗ trợ tìm kiếm và truy xuất thông tin.

Bài toán và pipeline đề xuất

Khóa luận tập trung vào pipeline phát hiện và nhận dạng văn bản biển hiệu trong video, bao gồm các mô-đun chính:

- **Trích xuất khung hình và tiền xử lý:** lấy mẫu khung hình từ video, hiệu chỉnh cơ bản (nếu cần) nhằm giảm nhiễu, mờ chuyển động và sai lệch ánh sáng.
- **Phát hiện biển hiệu (Signboard Detection):** xác định vùng chứa biển hiệu trong khung hình (dạng hộp bao hoặc đa giác), làm vùng quan tâm (ROI).

- **Phát hiện văn bản (Text Detection):** trong ROI biển hiệu, phát hiện vùng văn bản (theo hộp thẳng hoặc hộp xoay) để tăng độ chính xác.
- **Nhận dạng văn bản (Text Recognition):** nhận dạng chuỗi ký tự tiếng Việt/đa ngôn ngữ từ các vùng văn bản phát hiện được.
- **Hậu xử lý và hợp nhất theo thời gian:** loại nhiễu, chuẩn hóa chuỗi, và hợp nhất kết quả giữa các khung hình liên tiếp (tracking/temporal voting) để ổn định đầu ra.
- **Tầng ứng dụng:** khai thác văn bản nhận dạng để tìm kiếm/truy xuất theo từ khóa hoặc danh mục; cung cấp nền tảng cho các tác vụ downstream.

Như vậy, **đầu vào (Input)** của hệ thống bao gồm (xem thêm hình minh họa Hình 1.1):

- **Video hành trình đầu vào:** video camera hành trình trên đường phố Việt Nam, chứa các cảnh có biển hiệu ở nhiều điều kiện (ban ngày/ban đêm, mưa/nắng, đông người/ít người, nhiều góc nhìn).
- **Truy vấn người dùng (tùy chọn):** từ khóa hoặc danh mục (ví dụ: *r̈ihà thuốc¨*, *ïtrường học¨*) phục vụ chức năng tìm kiếm và lọc kết quả.

Và **đầu ra (Output)** của hệ thống là:

- **Danh sách các phát hiện biển hiệu:** mỗi mục gồm thời điểm (timestamp), vị trí biển hiệu (bounding box/polygon) và nội dung văn bản nhận dạng.
- **Danh sách các đoạn video/khung hình liên quan:** các đoạn video ngắn hoặc khung hình đại diện chứa biển hiệu phù hợp với truy vấn, được sắp xếp theo mức độ liên quan.

Mặc dù đã có nhiều tiến bộ trong lĩnh vực phát hiện và nhận dạng văn bản, bài toán phát hiện và nhận dạng văn bản biển hiệu trong video hành trình vẫn tồn tại nhiều thách thức đáng kể:

1. **Biến thiên điều kiện chụp và chất lượng ảnh:** mờ do chuyển động, độ phân giải thấp, nhiễu nén video và thay đổi ánh sáng làm giảm khả năng phát hiện và nhận dạng.

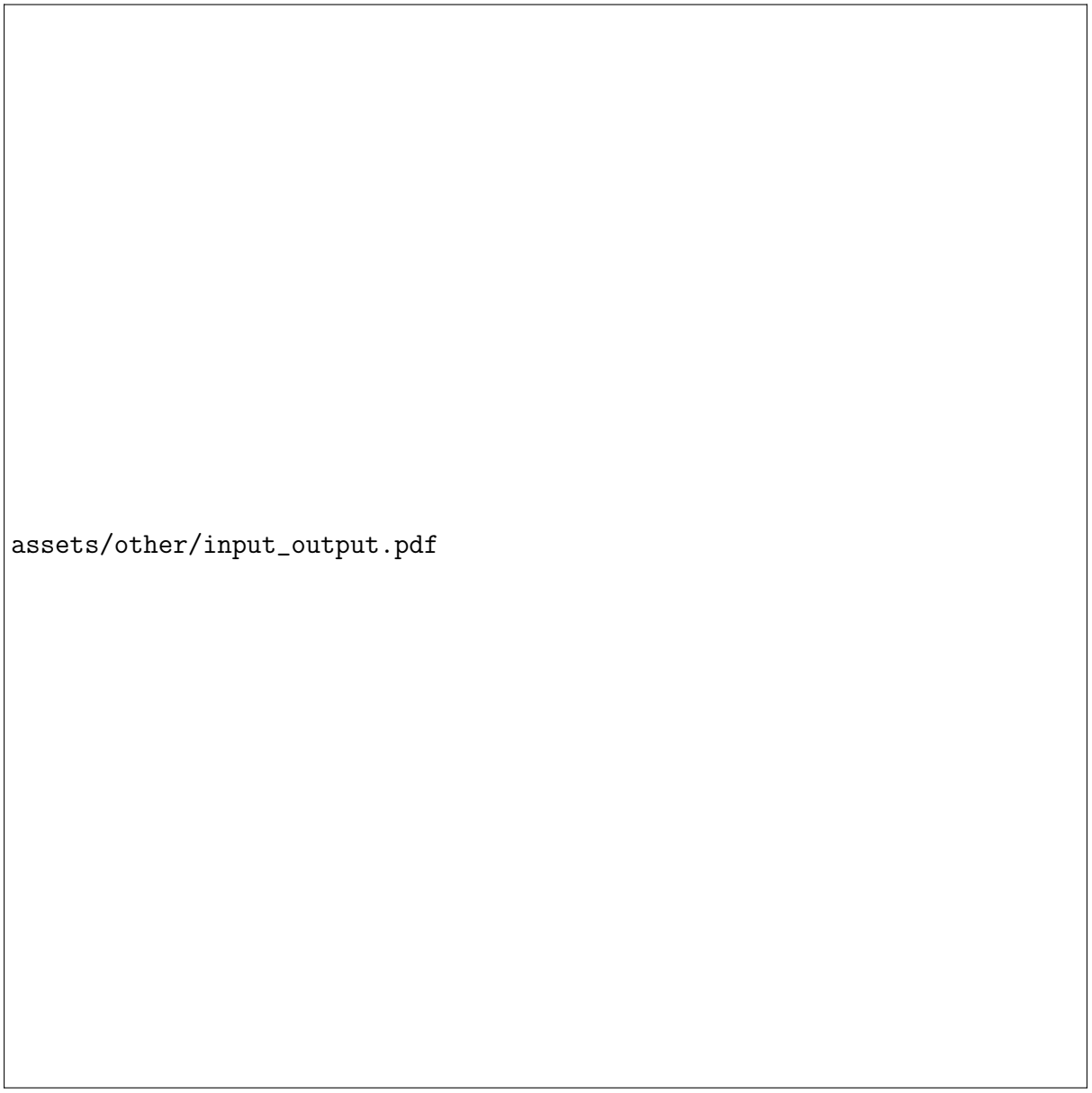

Figure 1.1: Minh họa đầu vào và đầu ra của hệ thống. Hệ thống tiếp nhận video (và/hoặc truy vấn) và trả về danh sách biển hiệu cùng nội dung văn bản nhận dạng, cũng như các đoạn video/khung hình liên quan.

2. **Đa dạng hình dạng biển hiệu và bố cục văn bản:** biển hiệu có thể cong, nghiêng, bị che khuất, nhiều lớp thông tin (logo, biểu tượng, chữ nghệ thuật), gây khó khăn cho cả detection và recognition.

3. **Đặc thù tiếng Việt và đa ngôn ngữ:** dấu tiếng Việt làm tăng độ phức tạp của bộ ký tự và dễ gây nhầm lẫn; thực tế biển hiệu có thể pha trộn Việt–Anh hoặc ký tự đặc biệt.

Từ những thách thức đã nêu, khóa luận này đặt ra mục tiêu phát triển một **pipeline phát hiện và nhận dạng văn bản biển hiệu trong video hành trình**, có khả năng:

- Xác định chính xác vùng biển hiệu trong khung hình video và phát hiện vùng văn bản tương ứng;
- Nhận dạng văn bản trên biển hiệu ổn định theo thời gian và hỗ trợ khai thác thông tin cho tác vụ tìm kiếm/truy xuất.

1.2 Mục tiêu và phạm vi

1.2.1 Mục tiêu

Trong khóa luận này, sinh viên đề ra các mục tiêu như sau:

- Khảo sát và tổng hợp các hướng tiếp cận tiên tiến để giải quyết bài toán STDR và bài toán văn bản biển hiệu.
- Mở rộng tập dữ liệu SignboardText [1] bằng cách bổ sung nhãn đối tượng biển hiệu (*signboard*).
- Cài đặt, thực nghiệm và đánh giá một số phương pháp hiện đại; phân tích ưu/nhược điểm của từng phương pháp.
- Xây dựng pipeline phát hiện và nhận dạng văn bản trên biển hiệu trong video.

- Phát triển ứng dụng minh họa khai thác thông tin văn bản từ biển hiệu, chẳng hạn như xác định loại hình cơ sở (cửa hàng, nhà hàng, trường học, bệnh viện...), nhằm hỗ trợ tìm kiếm và truy xuất theo từ khóa hoặc danh mục; đồng thời cung cấp nền tảng cho các tác vụ downstream như gợi ý địa điểm hoặc phân loại dịch vụ.

1.2.2 Phạm vi

Trong khóa luận này, nhóm sinh viên tập trung hoàn thành các công việc sau:

- Tìm hiểu tổng quan, thách thức và cơ sở lý thuyết của các phương pháp phát hiện biển hiệu, phát hiện và nhận dạng văn bản trong ảnh đời thường.
- Mở rộng tập dữ liệu SignboardText [1] bằng cách bổ sung nhãn biển hiệu; đồng thời thu thập thêm dữ liệu video hành trình từ camera hành trình trên đường phố Việt Nam.
- Cài đặt, thực nghiệm và đánh giá một số phương pháp hiện đại trên tập dữ liệu đã mở rộng; phân tích ưu/nhược điểm của từng phương pháp.
- Xây dựng pipeline phát hiện và nhận dạng văn bản trên biển hiệu trong video trên đường phố Việt Nam.
- Phát triển ứng dụng minh họa trên nền tảng web, cho phép khai thác thông tin văn bản từ biển hiệu nhằm hỗ trợ tìm kiếm, truy xuất thông tin và cung cấp dữ liệu đầu vào cho các hệ thống thông minh khác.

1.3 Đóng góp của khóa luận

Các đóng góp chính của khóa luận bao gồm:

- **Mở rộng bộ dữ liệu:** bổ sung nhãn đối tượng biển hiệu cho SignboardText [1] và xây dựng tập dữ liệu video hành trình phục vụ đánh giá pipeline.

- **Thực nghiệm và phân tích:** cài đặt và đánh giá nhiều phương pháp hiện đại cho các mô-đun (phát hiện biển hiệu/văn bản, nhận dạng văn bản), kèm phân tích ưu/nhược điểm theo bối cảnh tiếng Việt.
- **Pipeline và ứng dụng minh họa:** đề xuất pipeline STDR cho biển hiệu trong video và phát triển ứng dụng web hỗ trợ tìm kiếm/truy xuất theo từ khóa hoặc danh mục từ văn bản biển hiệu.

1.4 Cấu trúc khóa luận

Nội dung khóa luận được tổ chức như sau:

Chương 1: Tổng quan bài toán, bối cảnh, động lực, mục tiêu, phạm vi và đóng góp.

Chương 2: Cơ sở lý thuyết và các nghiên cứu liên quan đến phát hiện biển hiệu, phát hiện/nhận dạng văn bản và các kỹ thuật xử lý video.

Chương 3: Các phương pháp và pipeline đề xuất cho bài toán phát hiện và nhận dạng văn bản biển hiệu trong video, bao gồm mô tả kiến trúc hệ thống và mô-đun xử lý.

Chương 4: Thực nghiệm và đánh giá trên tập dữ liệu SignboardText mở rộng và dữ liệu video hành trình; phân tích kết quả và thảo luận.

Chương 5: Xây dựng ứng dụng minh họa và mô tả các chức năng khai thác thông tin văn bản biển hiệu.

Chương 6: Kết luận và hướng phát triển trong tương lai.

Chapter 2

CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN

2.1 Giới thiệu

2.1.1 Tổng quan và ý nghĩa thực tiễn của

2.1.2 Thách thức về tính

2.1.2.1 Bài toán)

Chapter 3

PHƯƠNG PHÁP

3.1 Hệ thống phát hiện và nhận dạng chữ trên biển hiệu

Chapter 4

THỰC NGHIỆM VÀ ĐÁNH GIÁ

4.1 Dữ liệu

4.2 Tiền xử lý

4.3 Tập câu truy vấn đánh giá

4.4 Độ đo đánh giá

4.5 Kết quả thực nghiệm

Chapter 5

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Kết luận

.....:

- aaaa.

5.2 Hướng phát triển

Để khắc phục những hạn chế trên và nâng cao hơn nữa tính hiệu quả, tính khả dụng và tính mở rộng của hệ thống, các hướng phát triển trong tương lai được đề xuất như sau:

Tối ưu hóa khả năng mở rộng dữ liệu:

- aaa
- aaa

Tăng cường khả năng tương tác và thích ứng với người dùng:

- Thiết kế giao diện người dùng aaaaaaaaaaaaa

Tích hợp truy vấn hình thức thoại (Spoken Query Integration): Phát triển hệ thống .....