

Text Detection and Recognition on Signboards in Vietnamese Street-View Videos

Nguyễn Đình Quân - 20521184, Nguyễn Hùng Phát - 22521074

December 21, 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

LỜI CẢM ƠN	i
Tóm tắt khóa luận	ii
Contents	iii
List of Figures	iv
List of Tables	v
1 TỔNG QUAN	1
1.1 Đặt vấn đề	1
1.2 Mục tiêu và phạm vi	7
1.2.1 Mục tiêu	7
1.2.2 Phạm vi	8
1.3 Đóng góp của khóa luận	8
1.4 Cấu trúc khóa luận	9
2 CƠ SỞ LÝ THUYẾT VÀ PHƯƠNG PHÁP TIẾP CẬN	10
2.1 Phát hiện đối tượng	10
2.1.1 Cơ sở và hướng tiếp cận chung	10
2.1.2 Phương pháp tiếp cận	11
2.2 Phát hiện và nhận dạng văn bản ngoại cảnh (Scene Text Detection and Recognition)	13

2.2.1	Phát hiện văn bản ngoại cảnh (Scene Text Detection)	13
2.2.1.1	Cơ sở và hướng tiếp cận chung	13
2.2.1.2	Phương pháp tiếp cận	14
2.2.2	Nhận dạng văn bản (Text Recognition)	15
2.2.2.1	Cơ sở và hướng tiếp cận chung	15
2.2.2.2	Phương pháp tiếp cận	15
2.2.3	End-to-End (End-to-End Text Recognition)	15
2.2.3.1	Cơ sở và hướng tiếp cận chung	15
2.2.3.2	Phương pháp tiếp cận	15
3	PHƯƠNG PHÁP	16
3.1	Hệ thống phát hiện và nhận dạng chữ trên biển hiệu	16
4	THỰC NGHIỆM VÀ ĐÁNH GIÁ	17
4.1	Dữ liệu	17
4.2	Tiền xử lý	17
4.3	Tập câu truy vấn đánh giá	17
4.4	Độ đo đánh giá	17
4.5	Kết quả thực nghiệm	17
5	KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	18
5.1	Kết luận	18
5.2	Hướng phát triển	18
References		20

List of Figures

1.1	Văn bản trong ảnh ngoại cảnh	2
1.2	Văn bản trên biển hiệu	3
1.3	Hình ảnh minh họa sự đa dạng về phông chữ, kích thước, chữ nghệ thuật, và đa ngôn ngữ [6]	4
1.4	Hình ảnh minh họa sự đa dạng về hình dạng, kích thước, vật liệu và vị trí của biển hiệu, dựa trên bộ dữ liệu SignboardText [6]	5
1.5	Hình ảnh minh họa đầu vào và đầu ra của bài toán. Bài toán nhận đầu vào là ảnh hoặc khung hình video và trả về danh sách các biển hiệu, trong đó mỗi biển hiệu được xác định bằng vùng chứa (bounding region) và nội dung văn bản tương ứng đã được nhận dạng.	6

List of Tables

Chapter 1

TỔNG QUAN

1.1 Đặt vấn đề

Phát hiện và nhận dạng văn bản trong ảnh ngoại cảnh (*Scene Text Detection and Recognition – STDR*) là một bài toán quan trọng trong thị giác máy tính, thu hút nhiều sự quan tâm nhờ tính ứng dụng rộng rãi như dịch tự động, hỗ trợ dẫn đường, hay phân tích biển báo giao thông. Với đầu vào là ảnh tĩnh hoặc các khung hình video, bài toán STDR hướng tới việc xác định vị trí xuất hiện và nội dung của văn bản (Hình 1.1).

Trong số các loại văn bản ngoại cảnh, **văn bản trên biển hiệu** (Hình 1.2) có ý nghĩa đặc biệt do thường chứa các thông tin quan trọng như *tên địa điểm*, *cơ sở kinh doanh* hoặc *loại hình dịch vụ*. Chính vì vậy, bài toán **phát hiện và nhận dạng văn bản trên biển hiệu** (*Text Detection and Recognition on Signboard*) trở thành một nhánh nghiên cứu quan trọng của STDR, với nhiều tiềm năng ứng dụng trong hệ thống dẫn đường thông minh, phân tích thông tin đô thị, và bổ sung thông tin ngữ nghĩa cho bản đồ số.

Tuy nhiên, bài toán phát hiện và nhận dạng văn bản trên biển hiệu đặt ra nhiều thách thức. Thách thức đầu tiên xuất phát từ đặc điểm của văn bản, như sự đa dạng về phông chữ, kích thước, hướng, bô cục; văn bản có thể bị nghiêng, cong, chồng chép hoặc hòa lẫn vào nền phức tạp, cùng với các phong cách thiết kế nghệ thuật và yếu tố đa ngôn ngữ (Hình 1.3). Đặc biệt đối với tiếng Việt, khó khăn còn gia tăng do hệ thống dấu thanh (sắc, huyền, hỏi, ngã, nặng) và các ký tự đặc biệt (ô, ê, ă, â, ơ, ư), làm tăng đáng kể tập ký tự cần nhận dạng và dễ gây nhầm lẫn giữa các chữ có hình dáng tương tự (ví dụ giữa



Hình 1.1: Văn bản trong ảnh ngoại cảnh

a, â, ă, á).

Thách thức thứ hai bắt nguồn từ đặc điểm của biển hiệu và bối cảnh môi trường xung quanh, biển hiệu đa dạng về hình dạng, kích thước, vật liệu và thường xuất hiện ở các vị trí phức tạp trong ảnh (Hình ??), chẳng hạn như bị che khuất một phần, chịu ảnh hưởng của phản xạ ánh sáng, hoặc nằm trong các bối cảnh đong đúc. Theo khảo sát các nghiên cứu hiện có, cho đến nay mới chỉ có một nghiên cứu [24] tập trung vào phát hiện biển hiệu trên đường phố Việt Nam, trong khi hướng tiếp cận kết hợp cả phát hiện đối tượng biển hiệu lẫn nhận dạng nội dung văn bản trên đó vẫn còn rất ít được khai thác.

Hơn nữa, khi mở rộng phạm vi từ ảnh tĩnh sang **video hành trình**, bài toán còn phải đổi mới với những thách thức đặc thù như hiện tượng mờ do chuyển động, chất lượng hình ảnh bị giới hạn bởi camera hành trình, cùng với sự biến đổi liên tục về điều kiện ánh sáng và góc quay. Những yếu tố này khiến nhiệm vụ phát hiện và nhận dạng văn



Hình 1.2: Văn bản trên biển hiệu

bản trong video trở nên phức tạp hơn nhiều so với trên ảnh đơn lẻ.

Từ những thách thức nêu trên, bài toán phát hiện và nhận dạng văn bản trên biển hiệu trong bối cảnh **video hành trình** có thể được định nghĩa một cách cụ thể như sau (hình ảnh minh họa trực quan tại Hình 1.5):

- **Đầu vào (Input):** Các hình ảnh hoặc khung hình thực tế được trích xuất từ video camera hành trình trên đường phố Việt Nam, chứa các cảnh có biển hiệu trong nhiều điều kiện khác nhau, bao gồm ban ngày/ban đêm, trời nắng/mưa và các góc nhìn đa dạng.
- **Đầu ra (Output):** Đối với mỗi hình ảnh (hoặc khung hình video) đầu vào, bài toán cần trả về hai thông tin chính:



Hình 1.3: Hình ảnh minh họa sự đa dạng về phông chữ, kích thước, chữ nghệ thuật, và đa ngôn ngữ [6]

- **Vị trí của biển hiệu:** Danh sách các vùng (bounding regions) xác định vùng chứa biển hiệu trong ảnh.
- **Thông tin văn bản trên từng biển hiệu:** Ứng với mỗi biển hiệu, cung cấp vị trí và nội dung văn bản đã được nhận dạng trên biển hiệu đó.

(Kết quả đầu ra có thể được trực quan hóa trực tiếp trên ảnh đầu vào hoặc tích hợp để xử lý liên tục cho luồng video.)

Trước những thách thức thực tế và dựa trên các kết quả nghiên cứu trước đây cho thấy rằng hướng nghiên cứu kết hợp (phát hiện biển hiệu và nhận dạng văn bản) vẫn còn ít được khai thác, khóa luận này đặt ra mục tiêu phát triển một **pipeline end-to-end** cho bài toán phát hiện và nhận dạng văn bản trên biển hiệu trong video được quay bởi camera hành trình trên đường phố. Pipeline hướng tới việc:

- Xác định vùng chứa biển hiệu (signboard detection) và vùng chứa văn bản bên trong mỗi biển hiệu (text detection) trong từng khung hình video.



Hình 1.4: Hình ảnh minh họa sự đa dạng về hình dạng, kích thước, vật liệu và vị trí của biển hiệu, dựa trên bộ dữ liệu SignboardText [6]

- Trích xuất và chuyển đổi nội dung văn bản từ các vùng văn bản đã phát hiện thành dạng văn bản có thể đọc được, hỗ trợ hai ngôn ngữ chính là tiếng Việt và tiếng Anh, hướng tới việc cung cấp thông tin đầu ra có ích cho các tác vụ truy xuất hoặc khai thác thông tin trong tương lai.

Để đạt được các mục tiêu trên, khóa luận sẽ tiến hành khảo sát, thực nghiệm so sánh và lựa chọn các phương pháp tiên tiến nay cho từng tác vụ con, đồng thời so sánh hai hướng tiếp cận chính cho bài toán text spotting. Các phương pháp cụ thể được xem xét bao gồm:

- **Phát hiện biển hiệu (Signboard Detection):** Các biến thể YOLO [25], DETR



Hình 1.5: Hình ảnh minh họa đầu vào và đầu ra của bài toán. Bài toán nhận đầu vào là ảnh hoặc khung hình video và trả về danh sách các biển hiệu, trong đó mỗi biển hiệu được xác định bằng vùng chứa (bounding region) và nội dung văn bản tương ứng đã được nhận dạng.

[3], RTDETR v2 [21], SegFormer [27], Mask2Former [5].

- **Phát hiện văn bản (Text Detection):** PANet [17], DBNet++ [15], TextPMs [29], FAST [4], KPN [31].
- **Nhận dạng văn bản (Text Recognition):** ViTSTR [1], PARSeq [2], CDistNet [34], SMTR [7], SVTRv2 [8]
- **Text Spotting (End-to-End):** TESTR [32], DeepSolo [28], UNITS [13], DNTextSpotter [23]

Trên cơ sở kết quả đánh giá và so sánh từ thực nghiệm cho từng tác vụ con, một pipeline end-to-end sẽ được xây dựng bằng cách lựa chọn phương pháp tối ưu cho mỗi tác vụ và xác định kiến trúc hiệu quả nhất cho giai đoạn xử lý văn bản thông qua so sánh hướng tiếp cận two-stage (tích hợp các phương pháp phát hiện và nhận dạng văn bản đã chọn) với các mô hình end-to-end tiên tiến.

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:

- Mở rộng và chuẩn bị tập dữ liệu ảnh tĩnh SignboardText [6] bằng cách bổ sung nhãn đối tượng biển hiệu (*signboard*), nhằm hỗ trợ đánh giá bài toán phát hiện biển hiệu.
- Thực nghiệm, so sánh và đánh giá một số phương pháp tiên tiến nay cho từng tác vụ con (phát hiện biển hiệu, phát hiện văn bản, nhận dạng văn bản) trên tập dữ liệu được chuẩn bị, từ đó rút ra ưu điểm, nhược điểm của từng phương pháp.
- Xây dựng một pipeline end-to-end cho bài toán phát hiện và nhận dạng văn bản trên biển hiệu trong video hành trình tại Việt Nam.

1.2.2 Phạm vi

Phạm vi của khóa luận được giới hạn nhằm đảm bảo tính tập trung và khả thi, bao gồm các công việc sau:

- Mở rộng tập dữ liệu tập trung vào việc bổ sung nhãn đối tượng biển hiệu (signboard annotation) cho tập dữ liệu ảnh tĩnh SignboardText hiện có. Dữ liệu video được thu thập chỉ nhằm mục đích minh họa và kiểm tra tính tổng quát của mô hình, với điều kiện chính là ban ngày. Các tình huống phức tạp (ban đêm, thời tiết xấu) không nằm trong phạm vi xem xét.
- Khảo sát và thực nghiệm được giới hạn trong một tập hợp các phương pháp tiên tiến cho các hướng tiếp cận phổ biến và hiệu quả hiện nay. Việc so sánh không bao quát toàn bộ các phương pháp trong lĩnh vực, mà tập trung vào những phương pháp phù hợp và khả thi với dữ liệu và mục tiêu của khóa luận.
- Pipeline end-to-end tập trung vào bài toán phát hiện và nhận dạng văn bản trên biển hiệu và hướng tới việc cung cấp thông tin đầu ra vị trí và nội dung văn bản, làm cơ sở cho các tác vụ truy xuất thông tin trong tương lai.

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 (signboard bounding box) cho tập dữ liệu ảnh tĩnh SignboardText [6], hỗ trợ thực nghiệm và đánh giá cho bài toán phát hiện biển hiệu. Đồng thời, thu thập một tập dữ liệu video hành trình thực tế để phục vụ minh họa và kiểm tra tính tổng quát.
- **Thực nghiệm và đánh giá:** Tiến hành cài đặt, thực nghiệm và so sánh một số phương pháp tiên tiến cho ba tác vụ thành phần: phát hiện biển hiệu, phát hiện văn bản và nhận dạng văn bản. Kết quả đánh giá đi kèm phân tích ưu/nhược điểm cụ thể trong bối cảnh dữ liệu tiếng Việt và cảnh quan đường phố.

- **Phát triển pipeline end-to-end:** Trên cơ sở kết quả thực nghiệm, phát triển một pipeline cho bài toán phát hiện và nhận dạng văn bản trên biển hiệu trong video hành trình trên đường phố Việt Nam. Pipeline hướng tới việc cung cấp đầu ra vị trí và nội dung văn bản, làm cơ sở cho các tác vụ truy xuất thông tin trong tương lai.

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 PHƯƠNG PHÁP TIẾP CẬN

2.1 Phát hiện đối tượng

2.1.1 Cơ sở và hướng tiếp cận chung

Phát hiện đối tượng (Object Detection) là một bài toán trong lĩnh vực Thị giác Máy tính (Computer Vision), đóng vai trò trung tâm trong nhiều ứng dụng thực tiễn như giám sát an ninh, lái xe tự động, và tương tác người máy. Khác với nhiệm vụ phân loại ảnh truyền thống vốn chỉ xác định loại đối tượng xuất hiện trong toàn bộ ảnh, phát hiện đối tượng yêu cầu mô hình không chỉ nhận diện đúng loại đối tượng mà còn xác định chính xác vị trí của chúng thông qua các hộp giới hạn (bounding boxes). Thách thức của bài toán này nằm ở việc phải xử lý đồng thời nhiều đối tượng với sự đa dạng lớn về kích thước, tư thế, góc nhìn, điều kiện ánh sáng và mức độ chồng lấn giữa các đối tượng.

Dựa trên tổng quan của [37], các phương pháp phát hiện đối tượng hiện đại có thể được phân loại thành ba hướng tiếp cận chính xét theo kiến trúc và quy trình xử lý:

- **Các phương pháp hai giai đoạn (Two-stage)** hoạt động dựa trên nguyên tắc tách biệt quá trình đề xuất vùng (region proposal) và phân loại. Nhóm này tiêu biểu bởi các mô hình thuộc họ R-CNN, chẳng hạn như R-CNN [10], Fast R-CNN [9] và Faster R-CNN [26]. Các phương pháp này thường đạt độ chính xác cao nhưng đòi hỏi chi phí tính toán lớn và tốc độ xử lý chậm.

- **Các phương pháp một giai đoạn (One-stage)** thực hiện trực tiếp việc dự đoán lớp và vị trí mà không có bước đề xuất vùng riêng biệt, với các đại diện nổi bật như YOLO [25], SSD [18] và RetinaNet [16]. Cách tiếp cận này giúp cân bằng tốt hơn giữa tốc độ và độ chính xác, phù hợp với các ứng dụng thời gian thực.
- **Các phương pháp dựa trên Transformer** gần đây tạo ra bước đột phá với kiến trúc end-to-end, loại bỏ sự phụ thuộc vào các thành phần được thiết kế thủ công (hand-crafted) như anchor và thuật toán Non-Maximum Suppression (NMS). Điển hình cho hướng đi này là mô hình DETR [3] và các biến thể tối ưu hóa tốc độ của nó.

2.1.2 Phương pháp tiếp cận

Trong bối cảnh của khóa luận này, bài toán phát hiện biển hiệu đòi hỏi sự cân bằng giữa tốc độ xử lý, độ chính xác và khả năng xử lý các đối tượng có hướng (oriented objects). Vì vậy, khóa luận tập trung lựa chọn và đánh giá một số phương pháp tiên tiến hiện nay, tiêu biểu cho các hướng tiếp cận khác nhau, dựa trên mức độ phù hợp với các yêu cầu của bài toán.

- **YOLO (You Only Look Once) [25]**: YOLO là đại diện tiêu biểu cho hướng tiếp cận một giai đoạn (one-stage). Kiến trúc của YOLO dựa trên việc chia ảnh đầu vào thành một lưới (grid), mỗi ô lưới chịu trách nhiệm dự đoán đồng thời các bounding box và xác suất lớp. Cách tiếp cận trực tiếp này mang lại tốc độ suy luận rất cao, phù hợp cho các ứng dụng thời gian thực. Đặc biệt, biến thể YOLO-OBB (Oriented Bounding Box) mở rộng khả năng phát hiện vật thể xoay, là một lựa chọn rất phù hợp cho bài toán phát hiện biển hiệu.
- **DETR (DEtection TRansformer) [3]**: DETR là mô hình phát hiện đối tượng đầu tiên hoàn toàn dựa trên kiến trúc Transformer. Mô hình sử dụng một tập hợp cố định các "truy vấn" (object queries) để tương tác với đặc trưng hình ảnh và dự đoán trực tiếp một tập các bounding box, nhờ đó loại bỏ hoàn toàn nhu cầu về các thành phần thủ công như anchor boxes và NMS.

- **RTDETRv2 [33]:** RT-DETRv2 là phiên bản cải tiến từ RT-DETR, được đề xuất với mục tiêu tối ưu hóa hiệu suất thời gian thực (real-time performance) trong khi vẫn duy trì độ chính xác cao. Mô hình này giữ nguyên ưu điểm end-to-end của DETR, loại bỏ sự phụ thuộc vào NMS, và được tối ưu hóa thông qua các cơ chế như hybrid encoder cùng cơ chế lựa chọn truy vấn (query selection) nhằm cân bằng giữa độ chính xác và hiệu quả tính toán.

Bên cạnh các phương pháp phát hiện trực tiếp dựa trên bounding box, để mở rộng góc nhìn đánh giá, khóa luận xem xét một hướng tiếp cận gián tiếp thông qua bài toán phân đoạn ngữ nghĩa (semantic segmentation). Theo hướng tiếp cận này, đối tượng trước hết được phân đoạn ở mức điểm ảnh, từ đó suy ra vùng bao hình học của đối tượng, phục vụ cho bài toán phát hiện. Trong bối cảnh đó, theo tổng quan của [?], các kiến trúc dựa trên Transformer đã trở thành một hướng tiếp cận được quan tâm rộng rãi và ngày càng quan trọng trong bài toán phân đoạn ảnh, đặc biệt là phân đoạn ngữ nghĩa. Nhờ khả năng mô hình hóa ngữ cảnh toàn cục thông qua cơ chế self-attention, các mô hình này cho thấy hiệu quả nổi bật trong việc xử lý các kịch bản có cấu trúc phức tạp và sự đa dạng lớn về hình dạng đối tượng.

- **SegFormer [27]:** SegFormer là một kiến trúc phân đoạn ngữ nghĩa hiệu quả dựa trên Transformer, kết hợp giữa encoder Transformer phân cấp và decoder MLP nhẹ. Thiết kế này cho phép mô hình khai thác ngữ cảnh toàn cục thông qua self-attention, đồng thời duy trì hiệu quả tính toán cao nhờ cấu trúc decoder đơn giản. Nhờ đó, SegFormer đạt được sự cân bằng tốt giữa độ chính xác và tốc độ suy luận, phù hợp với các ứng dụng phân đoạn ngữ nghĩa trong bối cảnh thực tế.
- **Mask2Former [5]:** Mask2Former là một kiến trúc phân đoạn dựa trên Transformer theo hướng tiếp cận thống nhất (unified framework), có khả năng xử lý nhiều bài toán phân đoạn khác nhau như phân đoạn ngữ nghĩa (semantic segmentation), phân đoạn theo thể hiện (instance segmentation) và phân đoạn toàn cảnh (panoptic segmentation). Mask2Former áp dụng cơ chế masked attention, cho phép mô hình tập trung vào các vùng đối tượng tiềm năng thông qua các mask dự đoán, từ đó

cải thiện khả năng biểu diễn hình dạng và biên đối tượng. Cách tiếp cận này giúp Mask2Former thể hiện hiệu quả trên các trường hợp có đối tượng chồng lấn hoặc hình dạng phức tạp.

2.2 Phát hiện và nhận dạng văn bản ngoại cảnh (Scene Text Detection and Recognition)

2.2.1 Phát hiện văn bản ngoại cảnh (Scene Text Detection)

2.2.1.1 Cơ sở và hướng tiếp cận chung

Phát hiện văn bản (Text Detection) trong ảnh ngoại cảnh hướng tới mục tiêu xác định và khoanh vùng các khu vực chứa văn bản. Khác với các tác vụ phát hiện đối tượng truyền thống, phát hiện văn bản trong ảnh ngoại cảnh phải đổi mới với nhiều thách thức do sự đa dạng về hình dạng, kích thước, hướng và bố cục của văn bản, cũng như các trường hợp văn bản bị nghiêng, cong, chồng chéo hoặc mờ. Do đó, bài toán này đòi hỏi kết hợp các kỹ thuật phát hiện đối tượng với các phương pháp chuyên biệt cho văn bản nhằm xác định chính xác và hiệu quả các vùng chứa văn bản.

Dựa trên các nghiên cứu khảo sát gần đây [12, 22, 11] về phát hiện văn bản trong ảnh ngoại cảnh, các phương pháp tiên tiến hiện nay có thể được phân thành ba hướng tiếp cận chính: dựa trên hồi quy (regression-based), dựa trên phân đoạn (segmentation-based) và dựa trên thành phần liên thông (connected component-based).

- **Các phương pháp dựa trên hồi quy (Regression-based):** Hướng tiếp cận này giải quyết bài toán phát hiện văn bản tương tự như phát hiện đối tượng, bằng cách trực tiếp dự đoán tọa độ các vùng văn bản dưới dạng hộp chữ nhật hoặc đa giác. Nhờ kiến trúc tối ưu cho việc dự đoán tọa độ, các phương pháp này thường có tốc độ suy luận nhanh, phù hợp với các ứng dụng thời gian thực. Tuy nhiên, một hạn chế chung của hướng tiếp cận này là thường yêu cầu các bước hậu xử lý (post-processing) phức tạp, đồng thời gặp khó khăn khi xử lý văn bản cong hoặc có hình dạng phức tạp. Một số phương pháp tiên tiến theo hướng này bao gồm TextBoxes

[14], EAST [35], FCE-Net [36], ABCNet [19].

- **Các phương pháp dựa trên thành phần liên thông (Connected Component-based):** Các phương pháp này tập trung vào việc phát hiện và nhóm các thành phần ảnh có đặc trưng tương đồng (như màu sắc, kết cấu hoặc cường độ) để hình thành các vùng văn bản hoàn chỉnh. Hướng tiếp cận này đạt hiệu quả trong các trường hợp đơn giản nhưng thường kém hiệu quả khi gặp nền phức tạp, văn bản cong hoặc có hình dạng phức tạp. Đồng thời, việc nhóm các thành phần riêng lẻ cũng đòi hỏi các bước hậu xử lý phức tạp để tái cấu trúc văn bản. Một số phương pháp tiêu biểu trong nhóm này gồm TextSnake [20], DRRG [30].
- **Các phương pháp dựa trên phân đoạn (Segmentation-based):** Nhóm phương pháp này xem phát hiện văn bản như một bài toán phân đoạn mức pixel, trong đó mỗi pixel được phân loại là văn bản hoặc nền. Từ kết quả phân đoạn, các vùng văn bản được suy ra và phục hồi thông qua các bước hậu xử lý. Cách tiếp cận này cho phép xử lý hiệu quả các văn bản có hình dạng phi chuẩn, như văn bản cong hoặc nghiêng, nhưng thường tồn kém tính toán hơn. Một số phương pháp tiên tiến hiện nay với hướng tiếp cận này bao gồm PANet [17], DBNet++ [15], TextPMs [29], FAST [4] và KPN [31].

2.2.1.2 Phương pháp tiếp cận

Trong bối cảnh đầy thách thức của phát hiện văn bản trên biển hiệu thực tế, khóa luận lựa chọn một số phương pháp tiên tiến hiện nay nhằm đánh giá khả năng xử lý các văn bản có hình dạng đa dạng. Những phương pháp này có khả năng duy trì độ chính xác đồng thời giảm thiểu các bước hậu xử lý phức tạp. Việc đánh giá các mô hình không chỉ giúp so sánh hiệu quả giữa các giải pháp hiện nay, mà còn cung cấp cái nhìn tổng quan về những hướng tiếp cận khả thi cho bài toán.

- **PANet [17]:** PANet là một kiến trúc phân đoạn hiệu quả với hai thành phần chính gồm Feature Pyramid Enhancement Module (FPEM), chịu trách nhiệm tạo bản đồ đặc trưng đa tỷ lệ, và Feature Fusion Module (FFM), thực hiện tổng hợp các

đặc trưng này. Bằng cách áp dụng phương pháp pixel aggregation trên bản đồ đặc trưng cuối cùng, PANet có thể nhóm chính xác các pixel văn bản vào các thể hiện tương ứng, đạt hiệu quả cao nhờ quy trình phân đoạn có chi phí tính toán thấp.

- **DBNet [15]:** DBNet++ là phiên bản cải tiến của DBNet, tích hợp cơ chế differentiable binarization (DB) trực tiếp vào mạng phân đoạn để tạo mask văn bản chính xác và ổn định, giảm đáng kể các bước hậu xử lý. Đồng thời, Adaptive Scale Fusion (ASF) được áp dụng để hợp nhất các đặc trưng đa tỷ lệ, cải thiện khả năng xử lý các văn bản có kích thước khác nhau.
- **TextPMs [29]:**
- **FAST [4]:**
- **KPN [31]:**

2.2.2 Nhận dạng văn bản (Text Recognition)

2.2.2.1 Cơ sở và hướng tiếp cận chung

2.2.2.2 Phương pháp tiếp cận

2.2.3 End-to-End (End-to-End Text Recognition)

2.2.3.1 Cơ sở và hướng tiếp cận chung

2.2.3.2 Phương pháp tiếp cận

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 aaaaaaaaaaaaaaa

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

.....

References

- [1] Rowel Atienza. Vision transformer for fast and efficient scene text recognition. In *International conference on document analysis and recognition*, pages 319–334. Springer, 2021. [7](#)
- [2] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *European conference on computer vision*, pages 178–196. Springer, 2022. [7](#)
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. [5](#), [11](#)
- [4] Zhe Chen, Jiahao Wang, Wenhai Wang, Guo Chen, Enze Xie, Ping Luo, and Tong Lu. Fast: Faster arbitrarily-shaped text detector with minimalist kernel representation. *arXiv preprint arXiv:2111.02394*, 2021. [7](#), [14](#), [15](#)
- [5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. [7](#), [12](#)
- [6] T. Do, T. Tran, T. Nguyen, D.-D. Le, and T. D. Ngo. Signboardtext: Text detection and recognition in in-the-wild signboard images. *IEEE Access*, 12:62942–62957, 2024. [4](#), [5](#), [7](#), [8](#)

- [7] Yongkun Du, Zheneng Chen, Caiyan Jia, Xieping Gao, and Yu-Gang Jiang. Out of length text recognition with sub-string matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2798–2806, 2025. [7](#)
- [8] Yongkun Du, Zheneng Chen, Hongtao Xie, Caiyan Jia, and Yu-Gang Jiang. Svtv2: Ctc beats encoder-decoder models in scene text recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20147–20156, 2025. [7](#)
- [9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. [10](#)
- [10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. [10](#)
- [11] Xu Han, Junyu Gao, Chuang Yang, Yuan Yuan, and Qi Wang. Spotlight text detector: Spotlight on candidate regions like a camera. *IEEE Transactions on Multimedia*, 2024. [13](#)
- [12] JianJun Kang, Mayire Ibrayim, and Askar Hamdulla. Overview of scene text detection and recognition. In *Proceedings of the 14th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, pages 661–666, 2022. [13](#)
- [13] Taeho Kil, Seonghyeon Kim, Sukmin Seo, Yoonsik Kim, and Daehee Kim. Towards unified scene text spotting based on sequence generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15223–15232, 2023. [7](#)
- [14] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. [14](#)

- [15] Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):919–931, 2022. [7](#), [14](#), [15](#)
- [16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. [11](#)
- [17] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018. [7](#), [14](#)
- [18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. [11](#)
- [19] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9809–9818, 2020. [14](#)
- [20] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. Textsnake: A flexible representation for detecting text of arbitrary shapes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 20–36, 2018. [14](#)
- [21] Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer. *arXiv preprint arXiv:2407.17140*, 2024. [7](#)
- [22] Umapada Pal, Arnab Halder, Palaiahnakote Shivakumara, and Michael Blumenstein. A comprehensive review on text detection and recognition in scene images. *Artificial Intelligence and Applications*, 2(4):229–249, 2024. [13](#)

- [23] Qian Qiao, Yu Xie, Jun Gao, Tianxiang Wu, Shaoyao Huang, Jiaqing Fan, Ziqiang Cao, Zili Wang, and Yue Zhang. Dntextspotter: Arbitrary-shaped scene text spotting via improved denoising training. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10134–10143, 2024. 7
- [24] D. L. Quang, K. V. Sy, H. L. Viet, S. P. Bao, and H. B. Quang. Signboards detection from street-view image using convolutional neural network: A case study in vietnam. In *Proceedings of the RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 394–397, Ho Chi Minh City, Vietnam, 2022. 2
- [25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 5, 11
- [26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 10
- [27] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 7, 12
- [28] Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Tongliang Liu, Bo Du, and Dacheng Tao. Deepsolo: Let transformer decoder with explicit points solo for text spotting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19348–19357, 2023. 7
- [29] Shi-Xue Zhang, Xiaobin Zhu, Lei Chen, Jie-Bo Hou, and Xu-Cheng Yin. Arbitrary shape text detection via segmentation with probability maps. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):2736–2750, 2022. 7, 14, 15

- [30] Shi-Xue Zhang, Xiaobin Zhu, Jie-Bo Hou, Chang Liu, Chun Yang, Hongfa Wang, and Xu-Cheng Yin. Deep relational reasoning graph network for arbitrary shape text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9699–9708, 2020. [14](#)
- [31] Shi-Xue Zhang, Xiaobin Zhu, Jie-Bo Hou, Chun Yang, and Xu-Cheng Yin. Kernel proposal network for arbitrary shape text detection. *IEEE transactions on neural networks and learning systems*, 34(11):8731–8742, 2022. [7](#), [14](#), [15](#)
- [32] Xiang Zhang, Yongwen Su, Subarna Tripathi, and Zhuowen Tu. Text spotting transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9519–9528, 2022. [7](#)
- [33] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16965–16974, 2024. [12](#)
- [34] Tianlun Zheng, Zhenpeng Chen, Shancheng Fang, Hongtao Xie, and Yu-Gang Jiang. Cdistrnet: Perceiving multi-domain character distance for robust text recognition. *International Journal of Computer Vision*, 132(2):300–318, 2024. [7](#)
- [35] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017. [14](#)
- [36] Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, and Wayne Zhang. Fourier contour embedding for arbitrary-shaped text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3123–3131, 2021. [14](#)

- [37] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023.

10