NÂNG CAO HIỆU QUẢ PHÁT HIỆN VĂN BẢN TIẾNG VIỆT TRONG ẢNH NGOẠI CẢNH DỰA TRÊN CƠ CHẾ TẬP TRUNG NGỮ CẢNH

Nguyễn Thị Thanh Tân[*], Huỳnh Văn Huy[#], Ngô Quốc Tạo[+]
[*] Đại học Điện Lực
[#] Trường Đại học Lạc Hồng
[+] Viện Công nghệ Thông tin – Viện Hàn lâm KH&CN Việt Nam

***Abstract*—** Bài báo này đề xuất một giải pháp để nâng cao hiệu quả phát hiện văn bản tiếng Việt trong ảnh ngoại cảnh. Về cơ bản, phương pháp phát hiện văn bản ở đây được đề xuất dựa trên ý tưởng xây dựng cơ chế tập trung ngữ cảnh (context attention) để học các thuộc tính hình học khác nhau nhằm tái tạo lại biểu diễn đa giác của các vùng văn bản. Hiệu quả của phương pháp đã được kiểm nghiệm trên hai tập dữ liệu ngoại cảnh tiếng Việt, bao gồm: VinText và VnSceneText. Các kết quả thực nghiệm cho thấy phương pháp đề xuất có khả năng phát hiện được các văn bản tiếng Việt có hình dạng và kích thước bất kỳ với độ chính xác cao và ổn định. Cụ thể, phương pháp đạt Precision (độ chính xác), Recall (độ phủ), Hmean (độ trung bình điều hòa) trên tập VinText là (85.63%, 87.94%, 86.77%) và trên tập VnSceneText là (85.14%, 87.23%, 86.17%). Các kết quả thực nghiệm cho thấy đây là một hướng tiếp cận khả thi đối với việc phát hiện văn bản tiếng Việt trong ảnh ngoại cảnh.

***Keywords*—** Văn bản ngoại cảnh, ảnh ngoại cảnh, vùng văn bản, phân đoạn, phát hiện, đặc trưng, ánh xạ, độ chính xác, độ phủ, độ trung bình điều hòa, tích chập, scale, batch, batch normal.

I. MỞ ĐẦU

Ảnh văn bản ngoại cảnh (scene text image) là những hình ảnh chứa văn bản được chụp trong một môi trường tự nhiên (ngoại cảnh) mà không phải là văn bản trong tài liệu được quét hoặc trong môi trường số hóa. Văn bản này có thể xuất hiện trên các bề mặt khác nhau như biển báo, quảng cáo, bảng chỉ dẫn, bảng menu, tên đường, và nhiều đối tượng khác trong đời sống hàng ngày.

Phát hiện văn bản trong ảnh ngoại cảnh (scene text detection) hiện được ứng dụng rất phổ biến trong thực tế, điển hình như nhận dạng bảng hiệu, hộp sản phẩm, nhãn hàng, v.v giúp cải thiện hiểu biết của khách hàng về sản phẩm hoặc dịch vụ; Đọc và nhận biết nhãn, thông số kỹ thuật trên các linh kiện. Nhận dạng biển báo giao thông hỗ trợ người khiếm thị, xe tự hành, v.v.

So với bài toán OCR truyền thống, việc phát hiện phát hiện văn bản tiếng Việt trong ảnh ngoại cảnh thường có độ phức tạp lớn và phải đối mặt với nhiều thách thức do các nguyên nhân sau: i) Ảnh có độ phức tạp cao do chứa nhiều loại đối tượng và nhiễu khác nhau; ii) Văn bản trong ảnh thường không chuẩn (irregular) với nhiều hình dạng, kích thước, hướng, màu sắc và kiểu chữ khác nhau; iii) Không có sự đồng nhất giữa các hình ảnh do ảnh được chụp từ nhiều góc độ, khoảng cách, ánh sáng và độ phân giải khác nhau; iv) Văn bản có thể bị che khuất bởi các đối tượng khác trong ảnh hoặc bởi các yếu tố khác như mờ, tối, chói sáng hoặc nhiễu; v) Bài toán phát hiện văn bản ngoại cảnh thường phải xử lý một lượng lớn các hình ảnh hoặc video, điều này đòi hỏi các thuật toán phải có khả năng xử lý dữ liệu nhanh và hiệu quả.

Đối với việc phát hiện văn bản ngoại cảnh tiếng Việt, ngoài việc phải đối mặt với các thách thức của bài toán phát hiện văn bản trong ảnh ngoại cảnh nói chung, còn gặp phải nhiều khó khăn, thách thức do đặc điểm cấu trúc và ngôn ngữ, điển hình như:

♦ Phức tạp về ngữ pháp và cấu trúc: Cấu trúc tiếng Việt phức tạp hơn với việc sử dụng dấu thanh điều chỉnh ngữ điệu của từ ngữ, cũng như cách viết liền và cách viết tách giữa các từ ghép, điều này tạo thêm thách thức trong việc nhận dạng và phân biệt các từ và cụm từ. Bên cạnh đó, các tầng dấu thanh trong ảnh ngoại cảnh đôi khi được in ẩn rất ngẫu hứng, có thể gây ra vấn đề dính chữ và dính dòng trong văn bản, làm giảm độ chính xác trong phát hiện và nhận dạng.

♦ Cỡ chữ và hình thức biểu đạt: Sự đa dạng về cỡ chữ và phong cách thể hiện, nhất là trong các biển hiệu, quảng cáo, gây khó khăn khi phải nhận dạng từ trong một bức ảnh với nhiều style khác nhau.

♦Sự đồng nhất về mặt hình thức: Một số từ trong tiếng Việt có thể viết rất gần nhau với chỉ sự khác biệt nhỏ về dấu thanh, điều này cản trở việc nhận dạng chính xác 100% mà không cần đến ngữ cảnh.

Thực tế cho thấy để đảm bảo độ chính xác nhận dạng cần có thêm tích hợp các cơ chế xử lý riêng biệt phù hợp

Tác giả liên hệ: Nguyễn Thị Thanh Tân,
Email: tanntt@epu.edu.vn
Đến tòa soạn: 8/2023, chỉnh sửa: 11/2023, chấp nhận đăng: 12/2023.

với các đặc thù của tiếng Việt. Bài báo này đề xuất một phương pháp hiệu quả để nâng cao hiệu quả phát hiện văn bản ngoại cảnh tiếng Việt dựa trên học sâu. Trong đó kiến trúc mạng học sâu ở đây được chia thành 04 khối chính: Khối thứ nhất được thiết kế dưới dạng một kiến trúc backbone để trích chọn đặc trưng từ mỗi ảnh đầu vào. Khối thứ 2 có dạng một kiến trúc mạng kim tự tháp (Feature Pyramid Network - FPN) có nhiệm vụ hợp nhất các đặc trưng được trích chọn từ khối thứ nhất. Khối tập trung ngữ cảnh (khối thứ 3) bao gồm các lớp tích chập được xây dựng dựa trên cơ chế tập trung (attention), có khả năng tự động học các thông tin ngữ cảnh để lựa chọn các đặc trưng có tính đại diện tốt hơn. Các đầu ra thu được từ khối tập trung ngữ cảnh sẽ được truyền đến khối thứ 4 (phân vùng văn bản) để tái tạo hình ảnh các vùng văn bản được phát hiện.

Các đóng góp chính của chúng tôi trong bài báo này bao gồm:

Thứ nhất, chúng tôi đã đề xuất một kiến trúc mạng học sâu (end-to-end), tích hợp nhiều công đoạn trong một bước xử lý để trích chọn đặc trưng văn bản tiếng Việt trong ảnh ngoại cảnh. Kiến trúc này ít nhạy cảm với nhiễu cũng như hướng của văn bản và có khả năng thích nghi và giải quyết tốt đối với các văn bản có hình dạng bất kỳ bao gồm cả các văn bản cong (curve text).

Thứ hai, chúng tôi đã tích hợp cơ chế tập trung ngữ cảnh (contextual attention mechanism) cho phép mô hình học cách tạo ra một ma trận trọng số (được gọi là ma trận attention). Ma trận này chỉ ra mức độ quan trọng của mỗi pixel trong ảnh đối với việc phát hiện văn bản. Các vùng có giá trị attention cao hơn sẽ được mô hình tập trung xem xét. Điều này giúp giảm thời gian xử lý của thuật toán cũng như giảm ảnh hưởng của những yếu tố không quan trọng trong hình ảnh. Từ đó giúp cho các công đoạn phân vùng văn bản và hậu xử lý phía sau đơn giản và hiệu quả hơn.

Các nội dung còn lại của bài báo được trình bày như sau: Phần 2 trình bày tóm lược các hướng tiếp cận liên quan. Ý tưởng chính và các bước thực hiện của phương pháp đề xuất được mô tả chi tiết trong phần 3. Phần 4 trình bày quá trình kiểm thử và đánh giá hiệu quả của phương pháp. Một số kết luận và hướng phát triển được đề cập trong phần 5.

II. HƯỚNG TIẾP CẬN LIÊN QUAN

Như đã phân tích ở trên, do các vấn đề thách thức thực tế đối với bài toán phát hiện văn bản ngoại cảnh. Hầu hết các phương pháp được đề xuất trong vài năm trở lại đây đều theo hướng tiếp cận học sâu. Về cơ bản, có thể phân chia các phương pháp này thành ba hướng tiếp cận chính: Dựa trên hồi quy (bounding-box regression based), dựa trên việc phát hiện từng phần văn bản (part-based) và dựa trên phân vùng (segmentation-based).

Hướng tiếp cận thứ nhất (bounding-box regression based) sử dụng các mô hình hồi quy để dự đoán các hộp giới hạn (bounding-box) của các vùng văn bản. M. Liao và cộng sự [10] đã hiệu chỉnh các anchors và tỷ lệ của các nhân tích chập dựa trên SSD [11] để phát hiện văn bản. Trong [12]- [13], các tác giả đã áp dụng tứ giác hồi quy để phát hiện văn bản đa hướng. P. He và cộng sự [14] đã đề xuất một cơ chế tập trung (attention mechanism) để xác định sơ bộ các vùng văn bản. Trong [15], các tác giả đề xuất phương pháp RRD dựa trên việc tách rời phân loại và hồi quy. Trong [16], [17] tác giả đề xuất các phương pháp không sử dụng anchors (anchor-free), áp dụng hồi quy ở mức điểm ảnh (pixel) cho các vùng văn bản đa hướng. L. Xie và cộng sự [18] đã giới thiệu một kiến trúc mạng RPN (Region Proposal Network) để xử lý vấn đề chuẩn hóa (scale) trong phát hiện văn bản ngoại cảnh. Nhìn chung, các phương pháp dựa trên hồi quy thường sử dụng các thuật toán hậu xử lý đơn giản như NMS (Non-Maximum Suppression). Tuy nhiên, hầu hết các phương pháp này đều không đạt được độ chính xác kỳ vọng đối với các văn bản đầu vào có hình dạng bất thường (chẳng hạn hình dạng cong – covered text).

Trong hướng tiếp cận dựa trên phát hiện từng phần văn bản, B. Shi và cộng sự [19] đã đề xuất phương pháp phát hiện các hộp giới hạn (bounding boxes) của các đoạn văn bản và dự đoán các liên kết của chúng để xử lý các vùng văn bản dài (long text). J. Tang và cộng sự [20] đã đề xuất một thuật toán nhóm các thành phần nhận biết cá thể (instance-aware) để phân tách các vùng văn bản một cách hiệu quả hơn và cải thiện thuật toán liên kết để phù hợp với các vùng văn bản có hình dạng tùy ý. Các phương pháp này nhìn chung đạt được hiệu suất tốt trong việc phát hiện các dòng văn bản dài. Tuy nhiên, các thuật toán liên kết khá phức tạp với siêu tham số được tạo bằng tay, dẫn tới khó điều chỉnh thuật toán và khả năng thích nghi với sự đa dạng của dữ liệu đầu vào không đạt hiệu quả cao.

Các phương pháp dựa trên phân đoạn thường kết hợp dự đoán mức điểm ảnh và thuật toán hậu xử lý để xác định các hộp giới hạn (bounding boxes). Zhang và cộng sự [21] đã phát hiện văn bản đa hướng bằng phân đoạn ngữ nghĩa và các thuật toán dựa trên MSER. Xue và cộng sự [22] đã sử dụng các đường bao để phân tách các vùng văn bản. Trong [23], [24] các tác giả phát hiện các vùng văn bản có hình dạng tùy ý theo hướng tiếp cận instance segmentation (phân đoạn cá thể) dựa trên Mask R-CNN [25]. W. Wang và cộng sự [26] đã đề xuất mở rộng chuẩn hóa lũy tiến (progressive scale) bằng cách phân đoạn các vùng văn bản với các nhân chuẩn hóa (scale kernels) khác nhau. Tian và cộng sự [27] đã đề xuất cơ chế nhúng pixel (pixel embedding) để gom nhóm các pixel từ kết quả phân đoạn. Trong [10] và [27], các tác giả đã đề xuất các thuật toán hậu xử lý mới cho kết quả phân đoạn, dẫn đến tốc độ suy luận (inference) thấp hơn.

Các phương pháp nhanh để phát hiện văn bản trong ảnh ngoại cảnh tập trung vào cả độ chính xác và tốc độ inference. Gần đây, Pengfei Wang và cộng sự [28] đã đề xuất phương pháp SAST để phát hiện văn bản ngoại cảnh với hình dạng bất kỳ. Phương pháp này sử dụng một framework huấn luyện đa nhiệm (multi-task) tập trung theo ngữ cảnh dựa trên mạng kiến trúc mạng FCN (Fully Convolutional Network) để học các thuộc tính hình học khác nhau nhằm tái tạo lại biểu diễn đa giác của các vùng văn bản. Các kết quả thực nghiệm cho thấy SAST có khả năng phát hiện văn bản ngoại cảnh có hình dạng bất kỳ với độ chính xác cao và tốc độ xử lý nhanh. Tuy nhiên, phương pháp này còn bị hạn chế trong việc phát hiện các

ký tự quá lớn hoặc quá nhỏ và văn bản trong ảnh có độ cong lớn. Zhu và cộng sự [29] đề xuất phương pháp FCE (Fourier Contour Embedding) để biểu diễn chu tuyến của các văn bản có hình dạng tùy ý dưới dạng ký các ký hiệu phổ nhỏ gọn và xây dựng kiến trúc mạng FCENet với một mạng backbone, mạng FPN (feature pyramid networks) và một mô hình hậu xử lý đơn giản với phép biến đổi Fourier ngược (IFT) và thuật toán NMS (Non-Maximum Suppression). Các kết quả thực nghiệm cho thấy phương pháp này đạt độ chính xác cao trên các tập dữ liệu CTW1500 và Total-Text. Tuy nhiên, phương pháp này gặp hạn chế trong trường hợp văn bản và nền không có sự tương phản rõ rệt hoặc văn bản có kích thước quá lớn. M. Liao và cộng sự [30] đã đề xuất phương pháp phát hiện văn bản có hình dạng bất kỳ dựa trên việc sử dụng thuật toán nhị phân hóa khả vi (Differentiable binarization) và cơ chế ASF (Adaptive Scale Fusion) để kết hợp thông tin từ các tỷ lệ (scale) khác nhau của ảnh một cách linh hoạt. Phương pháp này có khả năng giải quyết bài toán phát hiện các văn bản bị méo, các văn bản có hình dạng và kích thước bất kỳ một cách hiệu quả (đạt độ chính xác cao và thời gian xử lý nhanh). Tuy nhiên, hiệu suất của phương pháp có thể bị giảm đáng kể đối với các văn bản có nhiều nhiễu. Ngoài ra, hiệu quả của phương pháp còn bị phụ thuộc vào ngưỡng scale được chọn (phương pháp đạt hiệu suất cao nếu ngưỡng scale được chọn phù hợp với ảnh đầu vào).

Đối với bài toán phát hiện văn bản ngoại cảnh tiếng Việt, nhóm đã khảo sát và áp dụng thử nghiệm một số kiến trúc mạng học sâu tiên tiến như EAST [16], SAST [28], DB [29] (hiện đang là các kiến trúc mạng được đánh giá cao về mặt độ chính xác cũng như hiệu suất đối với các ngôn ngữ không dấu điển hình như tiếng Anh, Đức). Tuy nhiên, các kết quả thực nghiệm cho thấy độ chính xác của các phương pháp này không đảm bảo đối với việc phát văn bản tiếng Việt. Cụ thể, các phương pháp thường mắc lỗi phát hiện sót các tầng dấu hoặc bị phát hiện nhầm các từ /cụm từ mà có tầng dấu in không chuẩn (cách xa ký tự hoặc tầng dấu xuất hiện tại vị trí gây nhập nhằng giữa các từ/cụm từ) có nhiều tầng dấu như dấu mũ, dấu thanh trên, dấu thanh dưới. Hình 1 thể hiện kết quả thực tế của phương pháp DB[29] trên tập dữ liệu tiếng Việt. Trong đó, các tầng dấu bị phát hiện sai được đánh dấu bởi các hình elip màu đỏ.

Hình 1: Kết quả thực nghiệm của phương pháp DB[29]

Khảo sát thực tế cho thấy các kết quả nghiên cứu trong phát hiện và truy xuất văn bản tiếng Việt trong ảnh ngoại cảnh hiện còn rất hạn chế. Có một vài kết quả nghiên cứu tiêu biểu trong nhận dạng văn bản ngoại cảnh tiếng Việt được đã được công bố gần đây [31], [32]. Trong [31], các tác giả tập trung chính vào khâu nhận văn bản (text recognition). Nhóm tác giả đã đề xuất một cách tiếp cận kết hợp từ điển vào cả giai đoạn trainining và inference của hệ thống nhận dạng văn bản trong ảnh ngoại cảnh. Từ điển được sử dụng để tạo ra một danh sách các kết quả có thể và tìm kiếm kết quả phù hợp nhất với hình thức hiển thị của văn bản. Phương pháp này giúp xử lý tốt hơn các trường hợp mơ hồ gặp phải trong thực tế, và cải thiện hiệu suất tổng thể của các framework nhận dạng văn bản. Nhóm tác giả đã đóng góp cho cộng đồng tập dữ liệu văn bản ngoại cảnh VinText, phục vụ việc nghiên cứu và cải thiện chất lượng các thuật toán. Trong [32], nhóm tác giả trình bày kết quả thực nghiệm của một số phương pháp đã có như DBN, PMTD, PAN, FCENet, RCNN, SATRN, NRTR, RS trên tập dữ liệu VinText.

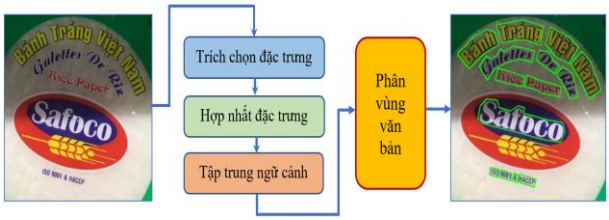
III. PHƯƠNG PHÁP PHÁT HIỆN VĂN BẢN NGOẠI CẢNH

Kế thừa từ các hướng tiếp cận đã có [16], [28], [29], chúng tôi đề xuất kiến trúc mạng học sâu với các khối chức năng chính: Trích chọn đặc trưng, hợp nhất đặc trưng, tập trung ngữ cảnh và phân vùng văn bản. Trong đó, các lớp mạng trong khối tập trung ngữ cảnh được thiết kế đặc biệt dựa trên cấu trúc ngôn ngữ tiếng Việt nhằm tăng cường các đặc trưng đại diện, giúp tăng độ chính xác của công đoạn phân vùng văn bản. Phương pháp đề xuất tập trung vào hai mục tiêu chính:

Thứ nhất, mô hình có khả năng phát hiện văn bản ngoại cảnh có hình dạng và kích thước bất kỳ với độ chính xác cao và tốc độ xử lý nhanh;

Thứ hai, mô hình có khả năng xử lý tốt đối với các tầng dấu mũ và dấu thanh điệu trong tiếng Việt.

Các bước cơ bản trong quy trình phát hiện văn bản ngoại cảnh của phương pháp đề xuất được mô tả cụ thể trên Hình 2. Từ mỗi ảnh đầu vào bất kỳ, trước tiên sẽ được đưa qua một mạng cơ sở (backbone) để tự động trích chọn các đặc trưng. Các đặc trưng được trích chọn sau đó sẽ được đưa qua mô hình hợp nhất. Mô hình hợp nhất được thiết kế đặc biệt nhằm tích hợp và biểu diễn các đặc trưng theo nhiều tỷ lệ khác nhau. Cơ chế xử lý này hướng tới mục tiêu giải quyết được các vùng văn bản có kích thước khác nhau.



Hình 2: Mô hình phát hiện văn bản ngoại cảnh

Các kết quả đầu ra thu được từ mô hình hợp nhất sẽ tiếp tục được chuyển tới mô hình tập trung ngữ cảnh (context attention mechanism) để tích hợp các phụ thuộc xa của các điểm ảnh nhằm thu được các đặc trưng có tính đại diện tốt hơn, giúp tăng độ chính xác của thuật toán

phân đoạn ở bước sau. Kết quả thu được từ bộ tập trung ngữ cảnh được sử dụng làm đầu vào cho công đoạn phân vùng văn bản (text instance segmentation). Bước phân vùng văn bản có nhiệm vụ định vị (xác định vị trí và hình dạng) của các vùng văn bản trong ảnh đầu vào. Phần sau đây chúng tôi sẽ mô tả các công đoạn thực hiện của phương pháp đề xuất một cách chi tiết và cụ thể hơn.

1. Trích chọn đặc trưng

Để trích chọn đặc trưng trên mỗi ảnh đầu vào chúng tôi sử dụng kiến trúc mạng backbone ResNet-50.. Kiến trúc này gồm 50 lớp (layer): Zero Padding (Lớp đệm), CONV (Lớp tích chập - Convolution), Max poolling, các Conv Block (Khối tích chập), các ID Block (Identity block – Khối định danh), Avg Pooling (Average Pooling) và Flattening. Quá trình thực hiện của mạng được chia thành 6 stage (công đoạn). Kiến trúc của stage 1 bao gồm 03 lớp tích chập (CONV), 64 filters (kích thước 7×7), stride (kích thước 2×2), chuẩn hóa BatchNorm, MaxPooling (3×3). Stage 2 bao gồm 01 khối Convolutiontal (Convolutiontal block) và 02 khối Identity (Identity block), 256 filters (kích thước 2×2), stride (kích thước 2×2). Stage 3 bao gồm 01 khối Convolutiontal và 03 khối Identity, 1024 filters (kích thước 2×2), stride (kích thước 2×2). Stage 4 bao gồm 01 khối Convolutiontal và 05 khối Identity, 1024 filters (kích thước 2×2), stride (kích thước 2×2). Đầu ra của stage 4 (FMS4) gồm 1024 ánh xạ đặc trưng, kích thước 48×64. Stage 5 bao gồm 01 khối Convolutiontal và 02 khối Identity, 2048 filters (kích thước 2×2), stride (kích thước 2×2). Stage 6 bao gồm 01 khối Convolutiontal và 02 khối Identity, 2048 filters (kích thước 2×2), stride (kích thước 2×2). Cấu trúc mỗi khối Conv gồm bốn lớp tích chập (conv) và một lớp kết hợp (concat).

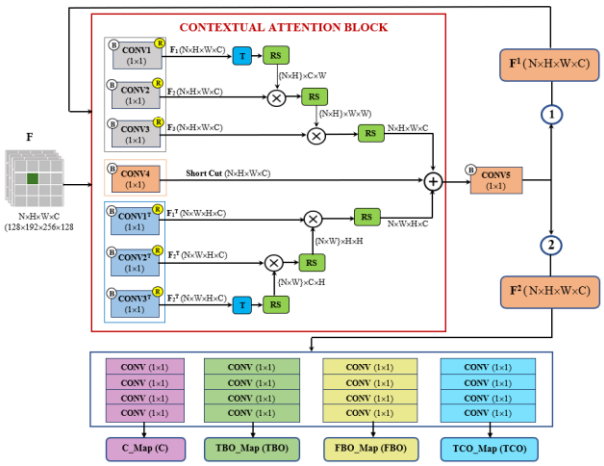
2. Hợp nhất đặc trưng

Trong kiến trúc mạng đề xuất, việc hợp nhất đặc trưng nhằm tích hợp và biểu diễn các đặc trưng theo nhiều tỷ lệ, giúp phát hiện các vùng văn bản có kích thước khác nhau. Về cơ bản, kiến trúc của mô hình được xây dựng dưới dạng một mạng FPN (Feature Pyramid Network). Mạng này được kiến trúc bởi hai luồng: Từ dưới lên (bottom-up pathway) và từ trên xuống (top-down pathway). Ba giá trị trong mỗi lớp mạng lần lượt tương ứng số kênh (channel), chiều cao và chiều rộng của các ánh xạ đặc trưng đầu ra. Luồng dưới lên (bottom-up pathway) đóng vai trò như một mạng encoder để trích chọn đặc trưng, càng lên cao độ phân giải (resolution) và kích thước đặc trưng càng giảm nhưng số filters (tương ứng với số ánh xạ đặc trưng đầu ra) và thông tin ngữ cảnh càng tăng. Luồng bottom-up trả về 128 ánh xạ đặc trưng (tương ứng với 128 kênh đầu ra).

Luồng từ trên xuống (top-down pathway) đóng vai trò như một mạng decoder có nhiệm vụ tái tạo lại đối tượng. Quá trình thực hiện của top-down pathway ngược lại với bottom-up pathway: càng đi xuống độ phân giải càng tăng, kích thước đặc trưng càng tăng, số filter (số ánh xạ đặc trưng đầu ra) càng giảm. Tập ánh xạ đặc trưng này tiếp tục được cung cấp cho mô hình tập trung ngữ cảnh (context attention) nhằm thu được các đặc trưng có tính đại diện tốt hơn, giúp tăng độ chính xác phát hiện/phân vùng văn bản ở công đoạn sau.

3. Cơ chế tập trung ngữ cảnh

Mô hình tập trung ngữ cảnh (contextual attention model - Hình 3) kế thừa cơ chế tự tập trung trong [28], [33]. Mục tiêu của mô hình tập trung ngữ cảnh nhằm tổng hợp các thông tin ngữ cảnh (mức cao) để tăng tính đại diện của các ánh xạ đặc trưng, giúp tăng độ chính xác của công đoạn phát hiện/phân vùng văn bản ở bước sau. Các bước thực hiện của mô hình tập trung ngữ cảnh được mô tả cụ thể trên Hình 3.



Hình 3: Mô hình tập trung ngữ cảnh (Context Attention)

Khối tập trung (Contextual Attention Block) tính toán các ma trận attention từ các đặc trưng ngang hoặc dọc và tích hợp thông tin ngữ cảnh qua phép nhân ma trận attention và đặc trưng gốc. Các phép biến đổi được sử dụng trong khối tập trung ngữ cảnh bao gồm:

Phép chuyển vị tập ánh xạ đặc trưng	T
Reshape tập ánh xạ đặc trưng	RS
Nhân hai tập ánh xạ đặc trưng	⊗
Kết hợp (concat) hai tập ánh xạ đặc trưng	⊕

Để giảm chi phí tính toán, cơ chế tập trung ngữ cảnh ở đây chỉ xem xét sự tương đồng của mỗi vị trí trong ánh xạ đặc trưng (feature map) với các vị trí khác trong cùng cột (ngữ cảnh theo chiều dọc) hoặc hàng (ngữ cảnh theo chiều ngang). Từ Sơ đồ thực hiện trên Hình 3 cho thấy khối tập trung ngữ cảnh được áp dụng hai lần liên tiếp nhằm mục đích lấy được cả các phụ thuộc gần và xa cho mỗi điểm ảnh. Cơ chế thực hiện cụ thể như sau:

Từ đầu vào là tập ánh xạ đặc trưng F (thu được từ mô hình hợp nhất), kích thước N×H×W×C, với N, H, W, C lần lượt là số lượng ánh xạ đặc trưng đầu vào, chiều cao, chiều rộng và số kênh của mỗi ánh xạ đặc. Thuật toán tập tập trung ngữ cảnh bao gồm các bước được mô tả cụ thể như sau:

Thuật toán 1: Tập trung ngữ cảnh cấu trúc tiếng Việt

1: Thực hiện 3 lớp tích chập (CONV1, CONV2, CONV3) 1×1 một cách song song để thu được 3 tập ánh xạ đặc trưng F_1, F_2 và F_3.

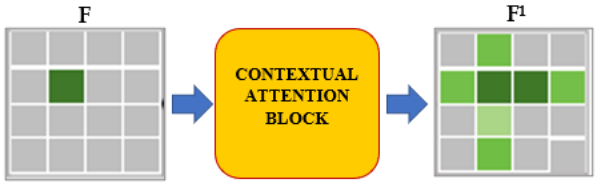
2: Chuyển vị F_1 và reshape kết quả về kích thước {N×H}×W×C;

3: Nhân F_2 với tập ánh chuyển vị thu được ở B2 và

reshape kết quả về kích thước {N×H}×W×W;
4: Nhân F_3 với tập ánh xạ tích thu được ở B3 và reshape kết quả về kích thước N×H×W×C.
5: Chuyển vị $F3^T$ và reshape kết quả về kích thước {N×W}×C×H;
6: Nhân $F2^T$ với tập ánh xạ chuyển vị thu được ở B5 và reshape kết quả về kích thước {N×W}×H×H;
7: Nhân $F1^T$ với tập ánh xạ tích thu được ở B6 và reshape kết quả về kích thước N×W×H×C.

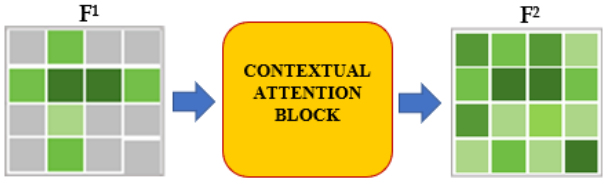
Bốn bước đầu tiên trong Thuật toán 1 được thực hiện để lấy thông tin ngữ cảnh tại mỗi điểm được xét theo chiều ngang và ba bước còn lại để lấy thông tin ngữ cảnh theo chiều dọc. Việc thu thập thông tin ngữ cảnh theo chiều dọc được thực hiện dựa trên các tập ánh xạ đặc trưng chuyển vị. Cụ thể, ba lớp tích chập ($CONV1^T$, $CONV2^T$, $CONV3^T$) được thực hiện song song để thu được lần lượt 03 tập ánh xạ chuyển vị tương ứng ($F1^T$, $F2^T$, $F3^T$) với shape N×W×H×C.

Trong mô hình tập trung, luồng short cut (short-cut path) được sử dụng trong kiến trúc với mục đích bảo toàn các đặc trưng cục bộ. Thao tác hợp (concat) tập ánh xạ ngữ cảnh theo chiều dọc, tập ánh xạ ngữ cảnh theo chiều ngang và tập đặc trưng short cut path để thu được tập đặc trưng đầu ra F^1 và giảm số kênh đầu ra bằng một lớp tích chập 1×1. Cơ chế này cho phép khối tập trung ngữ cảnh tổng hợp các thông tin ngữ cảnh theo cả hai hướng dọc và ngang cho mỗi điểm ảnh. Kết quả áp dụng tập trung ngữ cảnh lần 1 trên tập ánh xạ đặc trưng đầu vào F được thể hiện trên Hình 4.



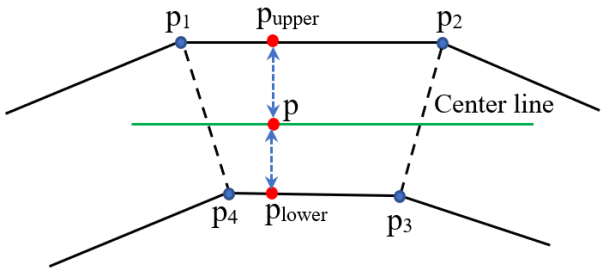
Hình 4: Áp dụng tập trung ngữ cảnh lần 1 trên tập F

Do các lớp tích chập (CONV1, CONV2, CONV3) và ($CONV1^T$, $CONV2^T$, $CONV3^T$) chia sẻ các trọng số nên việc tiếp tục áp dụng khối tập trung ngữ cảnh trên tập F^1 sẽ cho phép mỗi điểm ảnh lấy được các phụ thuộc xa hơn từ tất cả các điểm ảnh còn lại. Hình 5 cho thấy sau khi áp dụng khối tập trung ngữ cảnh lần 2, các phụ thuộc xa đối với mỗi điểm ảnh đã được tăng cường. Nói một cách khác là tập ánh xạ F^2 giàu thông tin ngữ cảnh hơn tập F^1.



Hình 5: Áp dụng khối tập trung ngữ cảnh trên tập F^1

Cơ chế xử lý này cũng giúp giảm bớt các vấn đề gây ra do khả năng tiếp nhận hạn chế khi xử lý các văn bản khó hơn, chẳng hạn như văn bản dài. Tập ánh xạ đặc trưng ngữ cảnh F^2 sẽ tiếp tục được sử dụng để dự đoán các đặc trưng (tri thức) mức cao của các vùng văn bản nhằm nâng cao độ chính xác của thuật toán phân vùng văn bản. Trong cách tiếp cận này, bốn loại đặc trưng mức cao được quan tâm chính bao gồm: Đường tâm (center line) của vùng văn bản các điểm biên (bounder point), các điểm góc của hình đa giác bao vùng văn bản, độ lệch (offset) của các điểm nằm trên đường tâm với các điểm biên. Việc trích chọn các đặc trưng mức cao ở đây được đảm nhiệm bởi bốn khối song song trong một mạng FCN - Fully Convolutional Network (được mô tả ở cuối Hình 3). Mỗi khối được cấu trúc với 4 lớp tích chập 1×1.



Hình 6: Độ lệch điểm p thuộc A và cặp điểm biên trên, dưới

Đầu ra của mỗi khối tương ứng với một loại ánh xạ đặc trưng mức cao cần xác định. Cụ thể, khối tích chập thứ nhất cho phép dự đoán ánh xạ đặc trưng biểu diễn đường tâm văn bản C_Map (ánh xạ đường tâm C - Hình 3). Khối tích chập thứ 2 cho phép dự đoán ánh xạ đặc trưng TBO_Map (ánh xạ TBO - Hình 3). Ánh xạ này xác định cặp điểm biên trên (p_{upper}), dưới (p_{lower}) của mỗi điểm p trong ánh xạ đường tâm C và ước lượng độ lệch (offset) giữa chúng.

Khối tích chập thứ 3 cho phép dự đoán ánh xạ FBO_Map (ánh xạ FBO - Hình 3). Ánh xạ này xác định bốn điểm góc của mỗi vùng văn bản (ví dụ bốn đỉnh p_1, p_2, p_3, p_4 trên Hình 6) và ước lượng độ lệch (offset) của bốn điểm này so với các điểm ảnh (pixel) trong ánh xạ đường biên C. Khối tích chập thứ 4 cho phép dự đoán ánh xạ TCO_Map (ánh xạ TCO), xác định độ lệch (offset) giữa các điểm ảnh (pixel) trong ánh xạ đường tâm C và điểm tâm của hình bao (bounding box) của khối văn bản.

4. Phân vùng văn bản

Phân vùng văn bản (text instance segmentation) là bước xử lý cuối cùng trong quy trình phát hiện văn bản. Đầu vào của bước này bao gồm các ánh xạ đặc trưng thu được từ mô hình tập trung ngữ cảnh ở bước trên.

Phân vùng văn bản là việc xác định vị trí các vùng văn bản (được thể hiện bởi các hình đa giác bao của chúng) trong ảnh đầu vào. Kế thừa ý tưởng trong [16], [28], [34], [35], thuật toán phân vùng văn bản được tiến hành như sau:

Thuật toán 2: Phân vùng văn bản

1: Xác định đường tâm (center line) của của vùng văn bản.
2: Lấy mẫu các điểm tâm (center point) trên đường center line.
3: Trích chọn các điểm biên của vùng văn bản.
4: Gom nhóm các điểm biên theo chiều kim đồng hồ để tạo thành các vùng văn bản.

Trong đó đường tâm văn bản được xem như hình ảnh thu nhỏ của vùng văn bản (Hình 7). Việc xác định đường tâm văn bản được thực hiện dựa trên ánh xạ đường tâm C

(đã được xây dựng từ mô hình tập trung ngữ cảnh ở trên).

Hình 7: Đường tâm văn bản

Phương pháp lấy mẫu các điểm tâm được thực hiện từ trái sang phải của đường tâm, các điểm tâm được lấy theo các khoảng cách đều nhau. Dựa trên các điểm tâm đã được lấy mẫu, bước xử lý tiếp theo sẽ tiến hành xác định các cặp điểm biên. Cuối cùng, tiến hành liên kết các điểm biên theo chiều kim đồng hồ, chúng ta sẽ thu được hình bao đầy đủ (vị trí và hình dạng) của các vùng văn bản.

IV. ĐÁNH GIÁ THỰC NGHIỆM

• Môi trường thực nghiệm

Nhóm tác giả sử dụng môi trường python 3 để cài đặt thuật toán và mô hình thử nghiệm. Cấu hình máy chạy thử Intel core i7-9700 CPU 4.7 GHz (Max Turbo Frequency), 32 GB RAM. Máy tính được trang bị Card đồ họa VGA Nvidia Tesla K80 12 GB × 2 GDDR5.

• Dữ liệu thực nghiệm

Để đánh giá hiệu quả của phương pháp, nhóm tác giả đã tiến hành thực nghiệm trên 02 tập dữ liệu văn bản ngoại cảnh tiếng Việt: VinText và VNSceneText. Thông tin cụ thể như sau:

Hình 8: Tập dữ liệu VinText

+ **VinText**. Tập dữ liệu văn bản ngoại cảnh tiếng Việt VinText được thu thập bởi Viện VinAI [31], với tổng số 1200 ảnh cho training và 300 ảnh cho testing (Hình 8). Các ảnh trong tập dữ liệu này được thu thập ở khu vực Hà Nội và các vùng lân cận, rất đa dạng, gồm ảnh chụp biển quảng cáo, biển tên đường, tên cửa hiệu, văn phòng, dòng chữ trên các phương tiện giao thông, v.v. Các ảnh trong tập dữ liệu này được gán nhãn mức từ.

Hình 9: Tập dữ liệu VNSceneText

+ **VNSceneText.** Đây là tập ảnh văn bản ngoại cảnh do nhóm tác giả tự thu trực tiếp trên đường phố khu vực Bà Rịa – Vũng Tàu và thành phố Hồ Chí Minh trong điều kiện hoàn toàn tự nhiên bằng các thiết bị smart phone (Iphone 7, Iphone 12, Oppo Reno 8). Tập dữ liệu bao gồm tổng số 3000 ảnh, trong đó 2400 ảnh dùng cho training và 600 ảnh dùng cho testing (Hình 9). Văn bản trong tập dữ liệu này rất đa dạng về chủng loại (ảnh chụp từ biển quảng cáo, biển chỉ dẫn giao thông, tên đường phố, biển hiệu trên các tòa nhà, phương tiện giao thông và ảnh chụp từ nhiều loại văn bản giấy tờ khác). Toàn bộ ảnh trong tập dữ liệu được gán nhãn mức từ theo định dạng Coco.

• Các độ đo đánh giá

Trong bài toán phát hiện đối tượng nói chung và bài toán phát hiện văn bản trong ảnh ngoại cảnh nói riêng, người ta thường sử dụng độ đo IoU để xác định một đối tượng hoặc một vùng văn bản có được phát hiện chính xác hay không. Độ đo IoU được định nghĩa bằng diện tích phần giao giữa vùng văn bản được phát hiện và vùng văn bản tương ứng đã được gán nhãn (lưu trong file ground truth) trên diện tích phần hợp của chúng.

Một vùng văn bản được coi là phát hiện đúng nếu độ đo IoU tương ứng của nó không nhỏ hơn một giá ngưỡng Γ cho trước.

$$\text{IoU} = \frac{\text{Diện tích phần giao}}{\text{Diện tích phần hợp}}$$

Để đánh giá hiệu quả của phương pháp, chúng tôi sử dụng các độ đo Precision, Recall và Hmean:

+ *Precision* (Độ chính xác): Là tỷ lệ giữa số lượng văn bản được phát hiện chính xác và tổng số văn bản được phát hiện:

$$Precision = \frac{T_P}{T_P + F_P}$$

+ *Recall* (Độ phủ): Là tỷ lệ giữa số lượng đối tượng được phát hiện chính xác và tổng số đối tượng thực tế:

$$Recall = \frac{T_P}{T_P + F_N}$$

+ *Hmean*: Là trung bình điều hòa giữa độ chính xác và độ phủ của các kết quả phát hiện:

$$Hmean = \frac{2 \times Precision \times Reall}{Precision + Reall}$$

Trong đó, *TP* (True Positive) là số vùng văn bản được phát hiện đúng trên ảnh; *FP* (False Positive) là số vùng văn bản bị phát hiện sai (không phải văn bản nhưng lại phát hiện là văn bản); *FN* là vùng văn bản không phát hiện được (bị bỏ sót).

Để đánh giá thời gian thực hiện của các thuật toán, chúng tôi sử dụng độ đo *FPS* (Frame Per Second), được định nghĩa là số khung hình (số ảnh) xử lý được trong một giây.

• Kết quả thực nghiệm

Quá trình thực nghiệm được tiến hành lần lượt trên từng tập dữ liệu. Đối với mỗi tập dữ liệu, trước tiên chúng tôi sẽ tiến hành huấn luyện mạng với phần dữ liệu training. Để đảm bảo tính ổn định của mô hình, từ phần dữ liệu training của mỗi mô hình, chúng tôi đã áp dụng các kỹ thuật làm giàu tập dữ liệu huấn luyện như thay đổi độ scale của ảnh (random_scale), xoay ảnh (random_rotate), làm mờ (random_blur), thay đổi độ tương phản ảnh (random_constrast), lật ngược ảnh (flip). Đối với mỗi ảnh trong tập training, chúng tôi thường sinh

thêm khoảng 10-15 ảnh để đưa vào huấn luyện.

Trong quá trình huấn luyện, các thành phần của của kiến trúc mạng (mạng backbone, mô hình hợp nhất, tập trung ngữ cảnh và phân vùng văn bản) được huấn luyện một các liên tiếp. Trong đó, riêng mạng backbone được huấn luyện theo hình thức transfer learning với pretrained weight từ ImageNet [38]. Các tham số cơ bản để huấn luyện các mô hình mạng được thiết lập như sau: Epoch: 5000; Batch size (trên 1 card): 16; Num_workers: 4; Optimizer: Adam, beta1: 0.9, beta2: 0.999; Learning_rate: 0.001; Regularizer: L2 normal.

Để có được các kết quả đánh giá một cách trực quan, chúng tôi đã so sánh kết quả thực hiện của phương pháp đề xuất với các phương pháp EAST [16], SAST [28], FCENet [29] và DB++ [30] như đã đề cập ở phần trên. Cách thức thực nghiệm được tiến hành như sau: Trước tiên, chúng tôi huấn luyện các mạng EAST, SAST, FCENet, DB++ trực tiếp trên tập VinText training và VNSceneText training theo cách thức transfer learning với pretrained weight từ [37]. Sau đó, chúng tôi sẽ tiến hành đánh giá và so sánh hiệu quả của các phương pháp trên các tập dữ liệu kiểm thử. Kết quả thực nghiệm của các phương pháp trên tập dữ liệu VinText được thể hiện cụ thể trên Bảng 1.

Bảng 1: Kết quả thực nghiệm trên tập dữ liệu VinText

Phương pháp	***Precision***	***Recall***	***Hmean***	***FPS***
EAST [16]	71.24	73.38	72.30	11
SAST [28]	83.12	85.03	84.06	8.5
FCENet [29]	83.23	84.57	83.89	9.2
DB++ [30]	84.05	83.90	83.97	8.3
Phương pháp đề xuất	85.63	87.94	86.77	7.9

Hình 10 thể hiện một số kết quả đầu ra của phương pháp đề xuất trên tập dữ liệu VinText.

Hình 10: Một số kết quả của phương pháp trên tập VinText

Kết quả thực nghiệm của các phương pháp trên tập dữ liệu VNSceneText được thể hiện cụ thể trên Bảng 2.

Bảng 2: Kết quả thực nghiệm trên tập dữ liệu VnSceneText

Phương pháp	Precision	Recall	Hmean	FPS
EAST [16]	70.73	72.16	71.44	10.03
SAST [28]	83.25	84.17	83.70	10.60
FCENet [29]	81.23	82.68	81.95	11.50
DB++ [30]	84.05	81.02	82.50	9.3
Phương pháp đề xuất	85.14	87.23	86.17	8.8

Hình 11 thể hiện một số kết quả của phương pháp đề xuất trên tập VNSceneText.

Hình 11: Một số kết quả trên tập VNSceneText

Các kết quả thực nghiệm trên cả hai tập dữ liệu cho thấy: Xét về độ chính xác, phương pháp đề xuất đạt độ chính xác cao hơn các phương pháp còn lại trên cả ba độ đo Precision, Recall và Hmean. Tuy nhiên, hiệu suất về thời gian của phương pháp giảm đi so với các phương pháp còn lại. Cụ thể, thời gian xử lý trung bình của phương pháp cỡ khoảng 0.12s ÷ 0.13s trên một ảnh. Thống kê từ các kết quả thực nghiệm cho thấy sai số của cả bốn phương pháp EAST [16], SAST [28], FCENet [29] và DB++ [30] đều tập trung chính trên các từ có cấu trúc nhiều tầng dấu. Các phương pháp SAST, FCENet, DB++ đạt hiệu suất về độ chính xác và thời gian xấp xỉ nhau, trong khi phương pháp EAST đạt hiệu suất vượt trội về mặt thời gian nhưng chưa xử lý tốt các trường hợp văn bản có hình dạng bất thường.

V. KẾT LUẬN

Trong bài báo này, chúng tôi đề xuất một mô hình hiệu quả để giải quyết bài toán phát hiện văn bản tiếng Việt trong ảnh ngoại cảnh dựa trên nền tảng học sâu. Phương pháp được đề xuất dựa trên ý tưởng sử dụng các kiến trúc mạng học sâu để học các thuộc tính hình học khác nhau nhằm tái tạo lại biểu diễn đa giác của các vùng văn bản. Kiến trúc mạng đề xuất gồm bốn thành phần chính: Mạng backbone (resnet-50) để trích chọn đặc trưng của ảnh đầu vào; Mô hình hợp nhất (fusion model), được xây dựng dưới dạng một mạng kim tự tháp đặc trưng (Feature Pyramid Network - FPN) nhằm tích hợp và biểu diễn các đặc trưng theo nhiều tỷ lệ từ đó giúp phát hiện các vùng văn bản có kích thước khác nhau; Mô hình tập trung ngữ cảnh (context attention), được cấu trúc dưới dạng một mạng tích chập sâu, được huấn luyện để lấy thông tin ngữ cảnh (cả các phụ thuộc gần và phụ thuộc xa) tại mỗi điểm ảnh nhằm thu được các đặc trưng có tính đại diện tốt hơn;

Mô hình phân vùng văn bản, được thực hiện theo cơ chế phân vùng ngữ nghĩa (semantic segmentation). Mô hình này có nhiệm vụ xác định vị trí và hình bao của tất cả các vùng văn bản có trong ảnh đầu vào dựa trên việc xác định đường tâm, điểm tâm và các điểm biên để gom nhóm chúng vào các vùng văn bản tương ứng. Điểm đặc biệt ở đây là trong quá trình gom cụm, thuật toán đã tích hợp cả các đặc trưng mức thấp (mức điểm ảnh) với các tri thức đối tượng mức cao nhằm thích nghi và giải quyết tốt đối với những văn bản có hình dạng phức tạp (cong, xiên, chữ to/nhỏ bất thường). Các kết quả thực nghiệm cho thấy đây là một hướng tiếp cận khả thi để giải quyết bài toán phát hiện văn bản tiếng Việt trong ảnh ngoại cảnh.

LỜI CẢM ƠN

Chúng tôi xin chân thành cảm ơn nhiệm vụ cao cấp, mã số NVCC02.01/24-24 đã hỗ trợ trong quá trình thực hiện nghiên cứu này.

REFERENCES

[1] Zobeir Raisi, Mohamed A. Naiel, Paul Fieguth, Steven Wardell, John Zelek, “Text Detection and Recognition in the Wild: A Review”, https://doi.org/10.48550/arXiv.2006.04305.

[2] S. Long, X. He, and C. Yao. Scene text detection and recognition: The deep learning era. Int. J. Comput. Vision, pages 1–24, 2020.

[3] S. M. Hanif and L. Prevost, “Text detection and localization in complex scene images using constrained adaboost algorithm,” in Proc. Int. Conf. on Doc. Anal. and Recognit., 2009, pp. 1–5.

[4] K. Wang, B. Babenko, and S. Belongie, “End-to-end scene text recognition,” in Proc. Int. Conf. on Comp. Vision, 2011, pp.1457–1464.

[5] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven,“PhotoOCR: Reading text in uncontrolled conditions,” in Proc. IEEE Int. Conf. on Comp. Vision, 2013, pp. 785–792

[6] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. Lim Tan, “Text flow: A unified text detection system in natural scene images,” in Proc. IEEE Int. Conf. on Comp. Vision, 2015, pp. 4651–4659.

[7] H. Cho, M. Sung, and B. Jun, “Canny text detector: Fast and robust scene text localization algorithm,” in Proc. IEEE Conf. on Comp. Vision and Pattern Recognit., 2016, pp. 3566–3573.

[8] B. Shi, X. Bai, and S. Belongie, “Detecting oriented text in natural images by linking segments,” in Proc. IEEE Conf. on Comp. Vision and Pattern Recognit., 2017, pp. 2550–2558.

[9] Y. Zhu, C. Yao, and X. Bai, “Scene text detection and recognition: Recent advances and future trends,” Frontiers of Comp. Sci., vol. 10, no. 1, pp. 19–36, 2016.

[10] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu. Textboxes: A fast text detector with a single deep neural network. In AAAI Conf. on Artificial Intelligence, 2017

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. E. Reed. SSD: single shot multibox detector. In European Conf. Comput. Vision, 2016.

[12] M. Liao, B. Shi, and X. Bai. Textboxes++: A single-shot oriented scene text detector. IEEE Trans. Image Processing, 27(8):3676–3690, 2018.

[13] Y. Liu and L. Jin. Deep matching prior network: Toward tighter multi-oriented text detection. In Proc. Conf. Comput. Vision Pattern Recognition, 2017.

[14] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li. Single shot text detector with regional attention. In Proc. Int. Conf. Comput. Vision, pages 3047–3055, 2017.

[15] M. Liao, Z. Zhu, B. Shi, G. Xia, and X. Bai. Rotation-sensitive regression for oriented scene text detection. In Proc. Conf. Comput. Vision Pattern Recognition, pages 5909–5918, 2018.

[16] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang. EAST: an efficient and accurate scene text detector. In Proc. Conf. Comput. Vision Pattern Recognition, 2017.

[17] W. He, X. Zhang, F. Yin, and C. Liu. Deep direct regression for multi-oriented scene text detection. In Proc. Int. Conf. Comput. Vision, 2017.

[18] L. Xie, Y. Liu, L. Jin, and Z. Xie. Derpn: Taking a further step toward more general object detection. In AAAI Conf. on Artificial Intelligence, volume 33, pages 9046–9053, 2019.

[19] B. Shi, X. Bai, and S. J. Belongie. Detecting oriented text in natural images by linking segments. In Proc. Conf. Comput. Vision Pattern Recognition, 2017.

[20] J. Tang, Z. Yang, Y. Wang, Q. Zheng, Y. Xu, and X. Bai. Seglink++: Detecting dense and arbitrary-shaped scene text by instanceaware component grouping. Pattern recognition, 96:106954, 2019. Qiao. Detecting text in natural image with connectionist text proposal network. In European Conf. Comput. Vision, 2016.

[21] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai. Multioriented text detection with fully convolutional networks. In Proc. Conf. Comput. Vision Pattern Recognition, 2016.

[22] C. Xue, S. Lu, and F. Zhan. Accurate scene text detection through border semantics awareness and bootstrapping. In European Conf. Comput. Vision, pages 355–372, 2018.

[23] M. Liao, P. Lyu, M. He, C. Yao, W. Wu, and X. Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. IEEE Trans. Pattern Anal. Mach. Intell., 2019.

[24] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In European Conf. Comput. Vision, pages 67–83, 2018.

[25] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. In ´ Proc. Int. Conf. Comput. Vision, pages 2961–2969, 2017.

[26] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, and S. Shao. Shape robust text detection with progressive scale expansion network. In Proc. Conf. Comput. Vision Pattern Recognition, pages 9336–9345, 2019.

[27] Z. Tian, M. Shu, P. Lyu, R. Li, C. Zhou, X. Shen, and J. Jia. Learning shape-aware embedding for scene text detection. In Proc. Conf. Comput. Vision Pattern Recognition, pages 4234–4243, 2019.

[28] Pengfei Wang, Chengquan Zhang, Fei Qi, Zuming Huang, Mengyi En, Junyu Han, Jingtuo Liu, Errui Ding, and Guangming Shi. A Single-Shot Arbitrarily-Shaped Text Detector based on Context Attended Multi-Task Learning. In Proceedings of the 27th ACM International Conference on Multimedia (MM ’19), October 21–25, 2019, Nice, France. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350988, 2019.

[29] Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, Wayne Zhang, "Fourier Contour Embedding for Arbitrary-Shaped Text Detection", CVPR, (2021), https://doi.org/10.48550/arXiv.2104.10442.

[30] M. Liao, Z. Zou, Z. Wan, C. Yao and X. Bai, "Real-Time Scene Text Detection With Differentiable Binarization and Adaptive Scale Fusion" in IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 45, no. 01, pp. 919-931, 2023.

[31] N. Nguyen et al., "Dictionary-guided Scene Text Recognition," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 7379-7388, doi: 10.1109/CVPR46437.2021.00730.

[32] N. T. Pham, V. D. Pham, Q. Nguyen-Van, B. H. Nguyen, D. N. Minh Dang and S. D. Nguyen, "Vietnamese Scene Text Detection and Recognition using Deep Learning: An Empirical Study," 2022 6th International Conference on Green Technology and Sustainable Development (GTSD), Nha Trang City, Vietnam, 2022, pp. 213-218, doi: 10.1109/GTSD54989.2022.9989248.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Adv. Neural Inf. Process. Syst. (NIPS). 5998–6008.

[34] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. 2018. TextSnake: A flexible representation for detecting text of arbitrary shapes. In Eur. Conf. Comp. Vis. (ECCV). 20–36.

[35] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. 2019. Shape Robust Text Detection With Progressive Scale Expansion Network. In IEEE Conf. Comp. Vis. Patt. Recognit. (CVPR). 9336–9345.

[36] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. 2015. ICDAR 2015 competition on robust reading. In Int. Conf. Doc. Anal. Recognit. (ICDAR). IEEE, 1156–1160.

[37] Chee Kheng Ch'ng and Chee Seng Chan. 2017. Total-Text: A comprehensive dataset for scene text detection and recognition. In Int. Conf. Doc. Anal. Recognit. (ICDAR), Vol. 1. IEEE, 935–942.
[38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In IEEE Conf. Comp. Vis. Patt. Recognit. (CVPR). IEEE, 248–255.

IMPROVING THE EFFICIENCY OF VIETNAMESE TEXT DETECTION IN SCENE IMAGES BASED ON CONTEXTUAL ATTENTION MECHANISM

Abstract—This paper proposes a solution to improve the efficiency of detecting Vietnamese text in scene images. The text detection method here is proposed based on the idea of building a contextual attention mechanism to learn different geometric properties to reconstruct the polygon representation of text regions. The effectiveness of the method has been experimented on two Vietnamese outdoor datasets, including VinText and VnSceneText. Experimental results show that the proposed method is capable of detecting Vietnamese texts of various shapes and sizes with high and stable accuracy. Specifically, the method achieves Precision, Recall, and Hmean on the VinText dataset as (85.63%, 87.94%, 86.77%) and on the VnSceneText dataset as (85.14%, 87.23%, 86.17%). This indicated that the proposed method is a feasible approach for detecting Vietnamese text in scene images.

Keywords—Scene text, scene images, text regions, segmentation, detection, features, mapping, accuracy, recall, harmonic mean, convolution, scale, batch, batch normalization.

Nguyễn Thị Thanh Tân tốt nghiệp đại học và nhận bằng Thạc sĩ lần lượt trong các năm 1999 và 2001 chuyên ngành Khoa học máy tính, ngành Công nghệ Thông tin tại Đại học Công nghệ - Đại học Quốc gia Hà Nội. Nhận bằng tiến sĩ chuyên ngành Khoa học máy tính tại Viện Hàn Lâm-Viện Khoa học và Công nghệ Việt Nam. Hiện nay tác giả công tác tại Trường Đại học Điện lực. Hướng nghiên cứu chính của tác giả bao gồm Trí tuệ nhân tạo và Thị giác máy tính, Phân tích và xử lý dữ liệu lớn. Tác giả đã công bố trên 40 bài báo tại các tạp chí có uy tín quốc gia và quốc tế.

Huỳnh Văn Huy tốt nghiệp đại học năm 2005 tại trường Đại học Khoa học Huế với chuyên ngành Tin học và nhận bằng Thạc sĩ năm 2012 chuyên ngành Công nghệ thông tin tại trường Đại học Lạc Hồng. Hiện nay tác giả đang là Nghiên cứu sinh ngành Khoa học Máy tính tại Trường Đại học Lạc Hồng.

Ngô Quốc Tạo đang công tác tại Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Ông nhận bằng PTS năm 1997 chuyên ngành "Đảm bảo toán học cho máy tính và các hệ thống tính toán", PGS Tin học năm 2002 và Nghiên cứu viên cao cấp năm 2018. Hiện tại ông đang nghiên cứu trong lĩnh vực Nhận dạng, Xử lý ảnh, Công nghệ tri thức, Trí tuệ nhân tạo. Ông đã tham gia công bố các bài báo có uy tín, tham gia hướng dẫn NCS cùng các đồng nghiệp