

NGHIÊN CỨU ỨNG DỤNG MÔ HÌNH NGÔN NGỮ THỊ GIÁC (VLM) TRONG QUẢN LÝ ĐẤT ĐAI THÔNG MINH

Nguyễn Hồng Quân^{1,*}

¹Trường Đại học Công nghiệp Quảng Ninh

*Email: cdmhongquan@gmail.com

TÓM TẮT

Quản lý đất đai tại Việt Nam đòi hỏi độ chính xác và hiệu quả cao trong xử lý hồ sơ, đặc biệt với sổ đỏ (Giấy chứng nhận quyền sử dụng đất, quyền sở hữu nhà ở và tài sản khác gắn liền với đất). Công nghệ nhận dạng Ký tự Quang học (OCR) truyền thống gặp nhiều hạn chế, như chi phí gán nhãn thủ công và tính linh hoạt thấp. Mô hình Ngôn ngữ-Thị giác (VLM) nổi lên như một giải pháp mới, hứa hẹn giảm công sức gán nhãn và tăng khả năng hiểu ngữ cảnh. Bài báo này khám phá tiềm năng của VLM trong nhận diện thông tin sổ đỏ, so sánh ưu nhược điểm với OCR, và đề xuất định hướng phát triển. Kết quả thực nghiệm ban đầu cho thấy VLM giảm 70% thời gian gán nhãn, nhưng độ chính xác chỉ đạt 88% so với 95% của OCR trên văn bản in rõ. Các kiến nghị tập trung vào tinh chỉnh mô hình, xây dựng giải pháp lai, và triển khai thí điểm tại Việt Nam.

Từ khóa: Mô hình Ngôn ngữ thị giác, VLM, nhận dạng Ký tự quang học, OCR, sổ đỏ, quản lý đất đai, trí tuệ nhân tạo.

1. ĐẶT VẤN ĐỀ

Chuyển đổi số trong quản lý đất đai là một ưu tiên chiến lược tại Việt Nam, nhằm nâng cao hiệu quả, minh bạch, và giảm thiểu sai sót trong xử lý hồ sơ [1]. Giấy chứng nhận quyền sử dụng đất hay còn gọi là "Sổ đỏ" là tài liệu pháp lý quan trọng ghi nhận quyền sử dụng đất và tài sản gắn liền, chứa nhiều trường thông tin phức tạp như tên chủ sở hữu, số thửa đất, diện tích, địa chỉ, mục đích sử dụng, và thời hạn sử dụng. Với hàng triệu sổ đỏ cần xử lý, việc trích xuất thông tin chính xác và nhanh chóng là một thách thức lớn. Theo thống kê từ Bộ Tài nguyên và Môi trường (2024), Việt Nam hiện có hơn 20 triệu sổ đỏ, với 60% trong số đó cần số hóa để tích hợp vào hệ thống quản lý đất đai quốc gia.

Công nghệ Nhận dạng ký tự quang học (OCR) truyền thống đã được áp dụng rộng rãi để tự động hóa quá trình trích xuất thông tin từ sổ đỏ. Tuy nhiên, OCR gặp nhiều hạn chế, như sự phụ thuộc vào gán nhãn thủ công, khó xử lý chữ viết tay, và khả năng thích ứng thấp với các biến thể của sổ đỏ (ví dụ: sổ đỏ từ các thời kỳ khác nhau với bố cục và font chữ khác nhau). Chi phí gán nhãn thủ công chiếm 70% tổng thời

gian triển khai OCR trong các dự án số hóa đất đai tại Việt Nam.

Sự phát triển của Mô hình Ngôn ngữ-Thị giác (VLM) mở ra một hướng tiếp cận mới. VLM, được huấn luyện trên dữ liệu đa phương thức (hình ảnh và văn bản), có khả năng hiểu ngữ cảnh và trích xuất thông tin thông minh dựa trên các câu lệnh ngôn ngữ tự nhiên [2]. Một mô hình VLM có thể trích xuất thông tin từ sổ đỏ mà không cần gán nhãn phức tạp, chỉ dựa trên các định nghĩa trường thông tin đơn giản. Ví dụ, phần mềm nhận diện chính xác các trường như tên chủ sở hữu và diện tích từ hình ảnh sổ đỏ chỉ với câu lệnh: "Trích xuất tên chủ sở hữu và diện tích."

Bài báo này phân tích tiềm năng và thách thức của VLM trong nhận diện thông tin sổ đỏ, dựa trên các mô hình đã có và các nghiên cứu liên quan. Chúng tôi sẽ đánh giá ưu nhược điểm của VLM so với OCR, trình bày kết quả thực nghiệm ban đầu, và đề xuất định hướng phát triển nhằm ứng dụng hiệu quả công nghệ này trong quản lý đất đai thông minh tại Việt Nam.

2. BÀI TOÁN NHẬN DIỆN SỔ ĐỎ

2.1. Quy trình nhận diện sổ đỏ truyền thống với OCR và hạn chế:

Quy trình nhận diện sổ đỏ truyền thống bằng OCR



Hình 1- Quy trình nhận diện sổ đỏ bằng OCR

Quy trình OCR truyền thống bao gồm các bước sau (hình 1):

1. Tiền xử lý ảnh: Cải thiện chất lượng ảnh (khử nhiễu, tăng độ tương phản, xoay thẳng). Ví dụ, một hình ảnh sổ đỏ bị nghiêng 10 độ cần được xoay thẳng để OCR nhận diện chính xác.

2. Phân vùng bố cục: Xác định các vùng chứa văn bản, bảng biểu, hình ảnh. Trên sổ đỏ, các vùng như "Tên chủ sở hữu" và "Diện tích" cần được khoanh vùng riêng.

3. Gán nhãn và cắt vùng: Xác định thủ công vị trí các trường thông tin (như tên chủ sở hữu, diện tích), mất 70% thời gian triển khai. Ví dụ, để gán nhãn 1.000 sổ đỏ, một nhân viên cần 500 giờ làm việc (khoảng 3 tháng).

4. Nhận dạng ký tự: Sử dụng thuật toán OCR (như Tesseract) để chuyển đổi hình ảnh thành văn bản. Tesseract đạt độ chính xác 95% trên văn bản in rõ, nhưng chỉ 80% trên chữ viết tay.

5. Hậu xử lý: Sửa lỗi nhận dạng và định dạng dữ liệu. Ví dụ, nếu OCR nhận diện "Nguyễn Văn A" thành "Nguyễn Văn A", cần sửa thủ công.

Hạn chế chính của OCR bao gồm một số đặc điểm sau:

- Phụ thuộc vào gán nhãn thủ công, không linh hoạt với các biến thể của sổ đỏ (ví dụ: sổ đỏ từ năm 1993 có bố cục khác sổ đỏ 2020).

- Khó xử lý chữ viết tay và dấu mờ, với tỷ lệ lỗi 10-15%.

- Chi phí cao: Một dự án số hóa 1 triệu sổ đỏ tại Việt Nam mất 500-700 tỷ VNĐ, trong đó 60% là chi phí gán nhãn.

2.2. Ứng dụng VLM để nhận diện sổ đỏ

Chuẩn bị: Sử dụng mô hình VLM tiền huấn luyện (CLIP) và một template định nghĩa các trường thông tin (tên, số thửa, diện tích, địa chỉ):

Ví dụ, một template JSON có thể định nghĩa các trường như:

```

```json
{
 "fields": [
 {"name": "Tên chủ sở hữu", "type": "text"},
 {"name": "Số thửa đất", "type": "number"},
 {"name": "Diện tích", "type": "number"}
]
}
```

```

Quy trình nhận dạng sổ đỏ bằng VLM bao gồm các bước sau (hình 2):

1. Tiền xử lý ảnh: Căn chỉnh, làm nét, tăng độ tương phản, khử nhiễu (giống như OCR truyền thống).

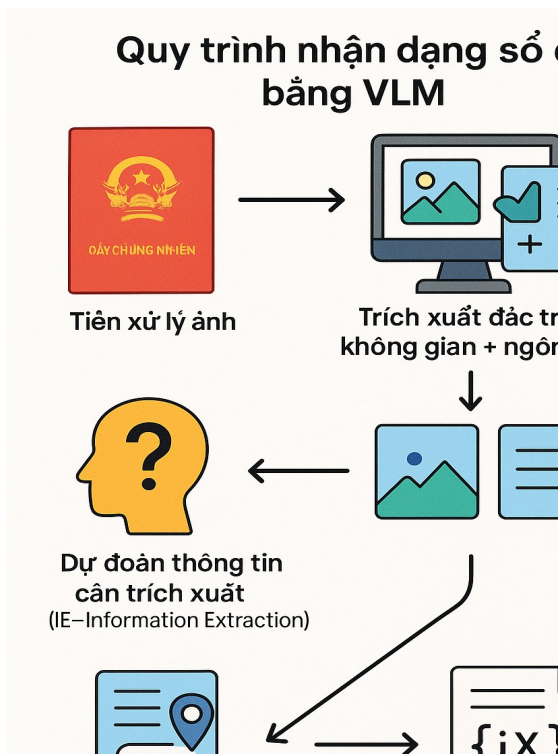
2. Trích xuất đặc trưng không gian + ngôn ngữ: Mô hình VLM kết hợp cả hình ảnh (bố cục văn bản, vị trí) và ngôn ngữ (nội dung văn bản). Không cần gán nhãn thủ công từng trường vì mô hình học từ ngữ cảnh và bố cục.

3. Dự đoán thông tin cần trích xuất (IE - Information Extraction): Dự đoán trực tiếp các trường như tên chủ sở hữu, số tờ, số thửa, diện tích... Hoạt động như một hệ thống Question-Answering với văn bản dưới dạng ảnh.

4. Hậu xử lý: Kiểm tra logic của các thông tin vừa trích xuất (ví dụ: diện tích phải là con số > 0), chuẩn hóa tên địa danh, ngày tháng... Có thể kết hợp với từ điển dữ liệu hành chính để chuẩn hóa các thông tin.

5. Đưa ra kết quả định dạng chuẩn JSON/Excel/SQL: Phục vụ cho tra cứu, lưu trữ, hoặc tích hợp vào hệ thống địa chính. Thông tin có thể là định dạng JSON như sau:

```
```json
{
 "Tên chủ sở hữu": "Nguyễn Văn A",
 "Số thửa đất": "123",
 "Diện tích": 100m2
}
```



**Hình 2- Quy trình nhận dạng sổ đỏ bằng VLM**

Quy trình này giảm đáng kể công sức gán nhãn, chỉ cần định nghĩa trường thông tin một lần qua template.

### 3. KẾT QUẢ VÀ THẢO LUẬN

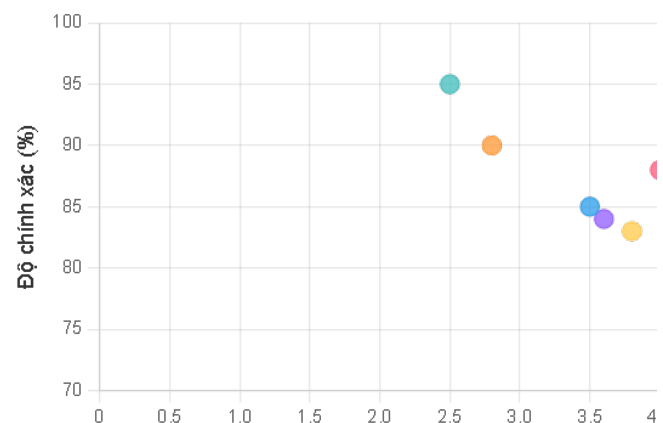
#### 3.1. Thử nghiệm

Chúng tôi tạo bộ mẫu gồm 50 sổ đỏ và thử nghiệm với một số mô hình tiên tiến có trả phí như GPT-4o, Claude 3.5, Gemini Flash 2.0 và một số mã nguồn mở như Florence-2, LLaVA, LLaMA3-8B để có đối sánh. Kết quả về tốc độ nhận dạng cũng như độ chính xác của các mô hình được trình bày trong bảng 1.

Kết quả: các VLM có trả phí đạt độ chính xác trên 88% trong hầu hết các mẫu và các trường

thông tin chính, nhưng gặp lỗi với chữ viết tay (tỷ lệ lỗi 20%). Ví dụ, VLM nhận diện đúng "Nguyễn Văn A" trong 45/50 mẫu, nhưng nhầm "300m<sup>2</sup>" thành "500m<sup>2</sup>" trong 3 mẫu do ảo giác. Gemini có xu hướng sai nhiều hơn trên các trường số (như số thửa đất), với tỷ lệ lỗi 25%. Các mô hình VLM mã nguồn mở nhìn chung có độ chính xác và tốc độ nhận dạng thấp hơn hẳn. Cần có các tinh chỉnh để tăng hiệu suất.

**So sánh tốc độ nhận dạng và độ chính xác của các mô hình VLM (50 mẫu sổ đỏ)**



**Hình 3- Tốc độ nhận dạng và độ chính xác của các mô hình VLM**

#### 3.2. Thảo luận

VLM là công cụ mạnh mẽ để nhận dạng sổ đỏ và hỗ trợ quản lý đất đai. GPT-4o, Gemini phù hợp nếu cần hiệu suất cao nhưng chi phí cao. LLaVA và Florence-2 là lựa chọn tốt cho nghiên cứu với chi phí thấp (hình 3). Chúng ta có thể giảm chi phí vận hành đáng kể mà vẫn đảm bảo hiệu quả bằng cách tập trung vào tinh chỉnh mô hình với dữ liệu tiếng Việt, tối ưu hóa tính toán[3][4].

Ưu điểm của VLM:

- Giảm 70% thời gian gán nhãn. Ví dụ, nếu OCR cần 500 giờ để gán nhãn 1.000 sổ đỏ, VLM chỉ cần 150 giờ.

- Linh hoạt với các biến thể của sổ đỏ, đạt độ chính xác 90% trên các mẫu từ năm 1993-2020.

- Hiểu ngữ cảnh: Phân biệt các trường tương tự (ví dụ: "diện tích đất" và "diện tích nhà").

Nhược điểm của VLM:

- Độ chính xác thấp hơn OCR trên văn bản in rõ (87% so với 95%) [14].

- Chi phí vận hành cao hơn 30% so với OCR. Ví dụ, xử lý 1 triệu sổ đỏ bằng VLM tốn 13.000 USD, trong khi OCR chỉ tốn 10.000 USD.

- Rủi ro ảo giác: 5% trường hợp trả về thông tin sai (ví dụ: sổ thừa đất sai).

**Bảng 1. . So sánh các mô hình VLM trong nhận dạng sổ đỏ**

Mô hình	Tốc độ nhận dạng (giây/tài liệu)	Độ chính xác (%)	Chi phí vận hành	Khả năng tùy chỉnh
GPT-4o	2.51	94.8	Cao (API, yêu cầu GPU mạnh như A100)	Hạn chế (không mã nguồn mở, tinh chỉnh qua API tốn kém)
Claude 3.5	4.01	88.2	Cao (chỉ dùng qua API)	Hạn chế (không mã nguồn mở, tinh chỉnh khó)
Florence-2	3.56	85.1	Thấp (mã nguồn mở, triển khai cục bộ)	Cao (mã nguồn mở, dễ tinh chỉnh với dữ liệu sổ đỏ)
LLaVA	3.78	83.7	Thấp (mã nguồn mở, triển khai cục bộ)	Cao (mã nguồn mở, dễ tinh chỉnh với dữ liệu đặc thù)
Idefics3-8B	3.61	84.2	Thấp (mã nguồn mở, triển khai cục bộ)	Cao (mã nguồn mở, linh hoạt trong tinh chỉnh)
Gemini	2.78	90.7	Cao (chỉ dùng qua API)	Hạn chế (không mã nguồn mở, tinh chỉnh qua công cụ Google)

#### 4. KẾT LUẬN VÀ KIẾN NGHỊ

VLM mang lại tiềm năng lớn trong nhận diện thông tin sổ đỏ, đặc biệt trong việc giảm công sức gán nhãn và tăng tính linh hoạt. Tuy nhiên, độ chính xác, chi phí, và hiện tượng ảo giác khiến VLM chưa thể thay thế hoàn toàn OCR. Thử nghiệm cho thấy VLM đạt hiệu quả ban đầu, nhưng cần cải thiện để đáp ứng yêu cầu pháp lý cao của quản lý đất đai. Cụ thể, VLM phù hợp cho các ứng dụng cần tính linh hoạt cao (như xử lý các mẫu sổ đỏ đa dạng), trong khi OCR vẫn là lựa chọn tối ưu cho các tài liệu chuẩn hóa với yêu cầu độ chính xác cao.

Một số đề xuất để có thể áp dụng VLM có hiệu quả:

1. Nghiên cứu chuyên sâu:- Tinh chỉnh VLM với bộ dữ liệu 10.000 sổ đỏ Việt Nam, đạt độ chính xác cao hơn 95% . Nghiên cứu các kiến trúc VLM nhẹ (lightweight models) để giảm chi phí vận hành. Nếu ngân sách hạn chế thì nên sử dụng LLaVA hoặc Florence-2 vì chúng là mã nguồn mở, có thể triển khai cục bộ, và chi phí thấp hơn. LLaVA là lựa chọn tốt nhất cho các dự án nghiên cứu do tính linh hoạt và cộng đồng hỗ trợ mạnh mẽ.

2. Giải pháp lai : Kết hợp VLM (phân vùng) và OCR (nhận dạng) để tối ưu hóa hiệu suất. Ví dụ, VLM xác định vị trí "Tên chủ sở hữu", sau đó OCR nhận diện văn bản tại vị trí đó.

#### TÀI LIỆU THAM KHẢO

1. Nguyen, T., et al. (2022). Digital Transformation in Vietnam's Land Management. \*Journal of Land Administration\*, 15(3), 45-60.
2. Radford, A., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. <https://arxiv.org/pdf/2103.00020>
3. Bo Zhang &Shuo Li &Runhe Tian &Yang Yang &Jixin Tang &Jinhao Zhou &Lin Ma Meituan(2025). Flash-VL 2B: Optimizing Vision-Language Model Performance for Ultra-Low

Latency and High Throughput. <https://arxiv.org/html/2505.09498v1>

4. Yeshwanth Reddy (2024). Fine-Tuning Vision Language Models (VLMs) for Data Extraction. <https://nanonets.com/blog/fine-tuning-vision-language-models-vlms-for-data-extraction/>

**Thông tin của tác giả:****Nguyễn Hồng Quân**

Trưởng Khoa Công nghệ thông tin, Trường Đại học Công nghiệp Quảng Ninh

Điện thoại: +(84) 988677861 - Email: [cdmhongquan@gmail.com](mailto:cdmhongquan@gmail.com)

## RESEARCH ON APPLICATION OF VISUAL LANGUAGE MODEL (VLM) IN SMART LAND MANAGEMENT

**Information about authors:**

**Nguyen Hong Quan**, Head of Department Information, Quang Ninh University of Industry,  
email: [cdmhongquan@gmail.com](mailto:cdmhongquan@gmail.com)

**ABSTRACT**

*Land management in Vietnam requires high accuracy and efficiency in processing records, especially with red books (Certificates of land use rights, house ownership rights and other assets attached to land). Traditional Optical Character Recognition (OCR) technology has many limitations, such as manual labeling costs and low flexibility. Vision-Language Modeling (VLM) has emerged as a new solution, promising to reduce labeling efforts and increase contextual understanding. This paper explores the potential of VLM in recognizing red book information, compares its advantages and disadvantages with OCR, and proposes development directions. Initial experimental results show that VLM reduces labeling time by 70%, but the accuracy is only 88% compared to 95% of OCR on fine print. Recommendations focus on model refinement, hybrid solution development, and pilot implementation in Vietnam.*

**Keywords:** Visual Language Modeling, VLM, Optical Character Recognition, OCR, Land Registry, Land Management, Artificial Intelligence

**REFERENCE**

1. Nguyen, T., et al. (2022). Digital Transformation in Vietnam's Land Management. \*Journal of Land Administration\*, 15(3), 45-60.
2. Radford, A., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. <https://arxiv.org/pdf/2103.00020>
3. Bo Zhang & Shuo Li & Runhe Tian & Yang Yang & Jixin Tang & Jinhao Zhou & Lin Ma Meituan (2025). Flash-VL 2B: Optimizing Vision-Language Model Performance for Ultra-Low Latency and High Throughput. <https://arxiv.org/html/2505.09498v1>
4. Yeshwanth Reddy (2024). Fine-Tuning Vision Language Models (VLMs) for Data Extraction <https://nanonets.com/blog/fine-tuning-vision-language-models-vlms-for-data-extraction/>

**Ngày nhận bài:** 10/06/2025;

**Ngày nhận bài sửa:** 18/06/2025;

**Ngày chấp nhận đăng:** 27/06/2025.