**OCR – Nhận dạng chữ viết**

**1. Khái niệm**

- OCR (Optical Character Regconition) là công nghệ nhận dạng ký tự quang học, có thể chuyển đổi hình ảnh văn bản thành định dạng văn bản máy tính có thể đọc được. OCR được sử dụng để trích xuất dữ liệu từ các tài liệu được quét, hình ảnh camera và tệp PDF chỉ có hình ảnh.

- Phần mềm OCR hoạt động bằng cách phân tích hình ảnh văn bản và xác định các chữ cái, từ và câu. Sau đó, nó ghép các chữ cái thành các từ và các từ thành câu.

**2. Ứng dụng**

- OCR có thể được sử dụng cho nhiều ứng dụng khác nhau, bao gồm những ứng dụng sau:

* Quét tài liệu đã in thành các phiên bản có thể chỉnh sửa bằng trình xử lý văn bản, như Microsoft Word hoặc Google Docs.
* Tự động nhập, trích xuất và xử lý dữ liệu.
* Giải mã tài liệu thành văn bản có thể đọc to cho người dùng khiếm thị hoặc mù.
* Lưu trữ thông tin lịch sử, chẳng hạn như báo, tạp chí hoặc danh bạ, ở dạng có thể tìm kiếm được.

**3. Phương pháp OCR**

**Bước 1**: **Thu nhận hình ảnh**

Một máy quét sẽ đọc tài liệu và chuyển đổi chúng thành dữ liệu nhị phân. Phần mềm OCR phân tích hình ảnh đã quét và phân loại vùng sáng làm nền và vùng tối làm văn bản.

**Bước 2** : **Tiền xử lý**

Trước tiên, phần mềm OCR sẽ làm sạch hình ảnh và loại bỏ các lỗi để chuẩn bị cho bước đọc. Sau đây là một số kỹ thuật làm sạch của phần mềm OCR:

* Chỉnh thẳng hoặc nghiêng nhẹ tài liệu đã quét để khắc phục lỗi về căn chỉnh trong quá trình quét.
* Khử nhiễu đốm hoặc loại bỏ mọi đốm ảnh kỹ thuật số hay làm mịn các viền của hình ảnh văn bản.
* Làm sạch đường viền khung và đường thẳng trong hình ảnh.
* Nhận dạng chữ viết cho công nghệ OCR đa ngôn ngữ.

**Bước 3** : **Nhận dạng văn bản**

- Hai loại thuật toán OCR mà phần mềm OCR sử dụng để nhận dạng văn bản được gọi là so khớp mẫu và trích xuất đặc điểm.

* ***Quy trình so khớp mẫu bao gồm các bước sau****:*

1. Hình ảnh ký tự được phân tách thành các hình dạng chữ.
2. Các hình dạng chữ được so sánh với các hình dạng chữ tương tự trong cơ sở dữ liệu.
3. Ký tự có hình dạng khớp nhất với hình dạng đầu vào được xác định.

So khớp mẫu chỉ hoạt động hiệu quả khi hình dạng chữ trong cơ sở dữ liệu có phông chữ và tỷ lệ tương tự với hình dạng chữ trong hình ảnh đầu vào. Phương thức này hoạt động tốt đối với hình ảnh quét từ tài liệu được đánh máy bằng phông chữ đã biết.

* ***Quy trình trích xuất đặc điểm bao gồm các bước sau****:*

1. Phân tách hình ảnh ký tự thành các đặc điểm như nét thẳng, nét vòng khép kín, hướng nét và giao điểm nét.
2. Tính toán các giá trị cho các đặc điểm này.
3. Sử dụng các giá trị này để tìm kết quả phù hợp nhất hoặc kết quả gần đúng nhất trong số các hình dạng chữ khác nhau được lưu trữ.

Trích xuất đặc điểm có thể hoạt động tốt đối với hình ảnh văn bản có chất lượng thấp hoặc bị nhiễu. Phương pháp này cũng có thể được sử dụng để nhận dạng ký tự từ các ngôn ngữ khác nhau.

**Bước 4** : **Hậu xử lý**

Sau khi phân tích, hệ thống sẽ chuyển đổi dữ liệu văn bản được trích xuất thành tệp trên máy tính. Một số hệ thống OCR có thể tạo các tệp PDF có chú thích bao gồm cả phiên bản trước và sau của tài liệu được quét.

**4. Thử nghiệm các công cụ**

**1. OCR Text Recognition Tool**

Ảnh có chứa văn bản, ảnh chụp màn hình, Phần mềm đa phương tiện, phần mềm

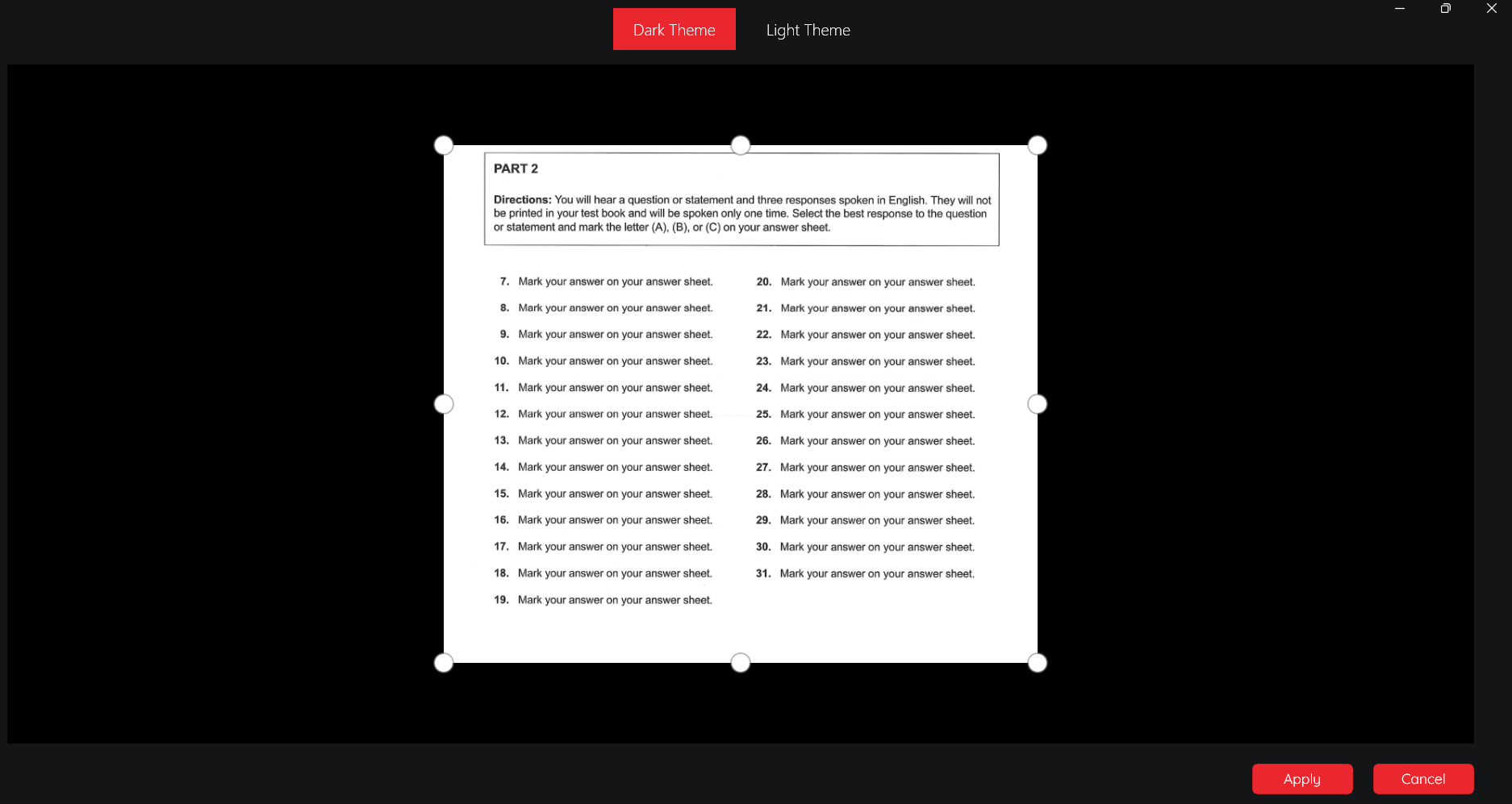
Mô tả được tạo tự động

***Giao diện***

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Phần mềm đa phương tiện

Mô tả được tạo tự động

***Thêm hình ảnh vào***



***Apply hình ảnh để xuất ra văn bản***

Ảnh có chứa văn bản, ảnh chụp màn hình, thiết kế

Mô tả được tạo tự động

***Thành quả***

**2. Convertio**

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Trang web

Mô tả được tạo tự động

***Giao diện***

Ảnh có chứa văn bản, bức thư, tài liệu, giấy

Mô tả được tạo tự động

***Hình ảnh đưa vào***

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Biểu tượng máy tính

Mô tả được tạo tự động



***Chọn hình ảnh muốn thêm vào***

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, Trang web

Mô tả được tạo tự động

***Giao diện sau khi đưa file hình ảnh vào***

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình

Mô tả được tạo tự động

***Giao diện sau khi nhận diện hình ảnh thành công***

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, số

Mô tả được tạo tự động

***Thành quả***

**3. SORA PDF**

Ảnh có chứa văn bản, phần mềm, Biểu tượng máy tính, Trang web

Mô tả được tạo tự động

***Giao diện***

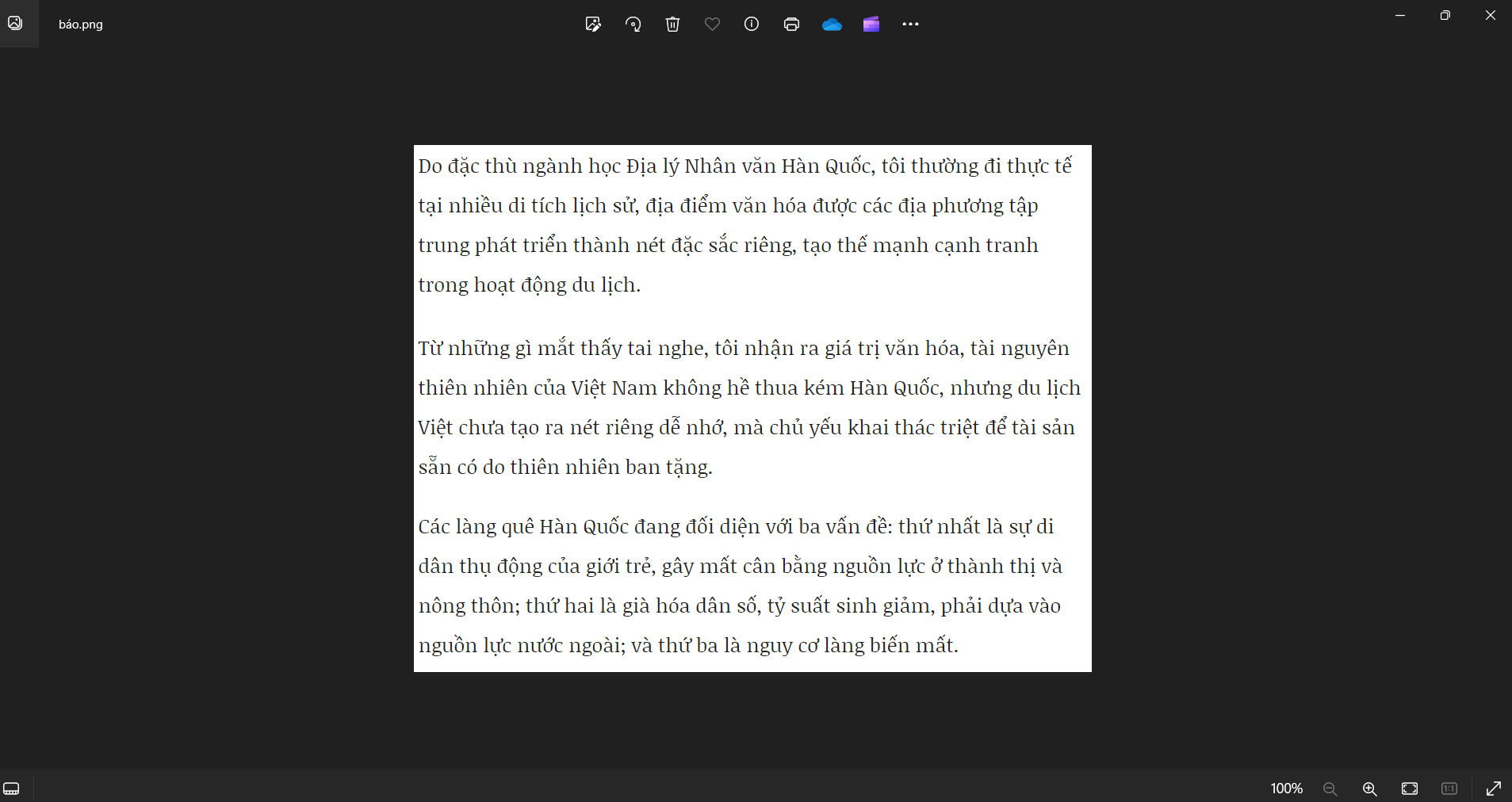


Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Biểu tượng máy tính

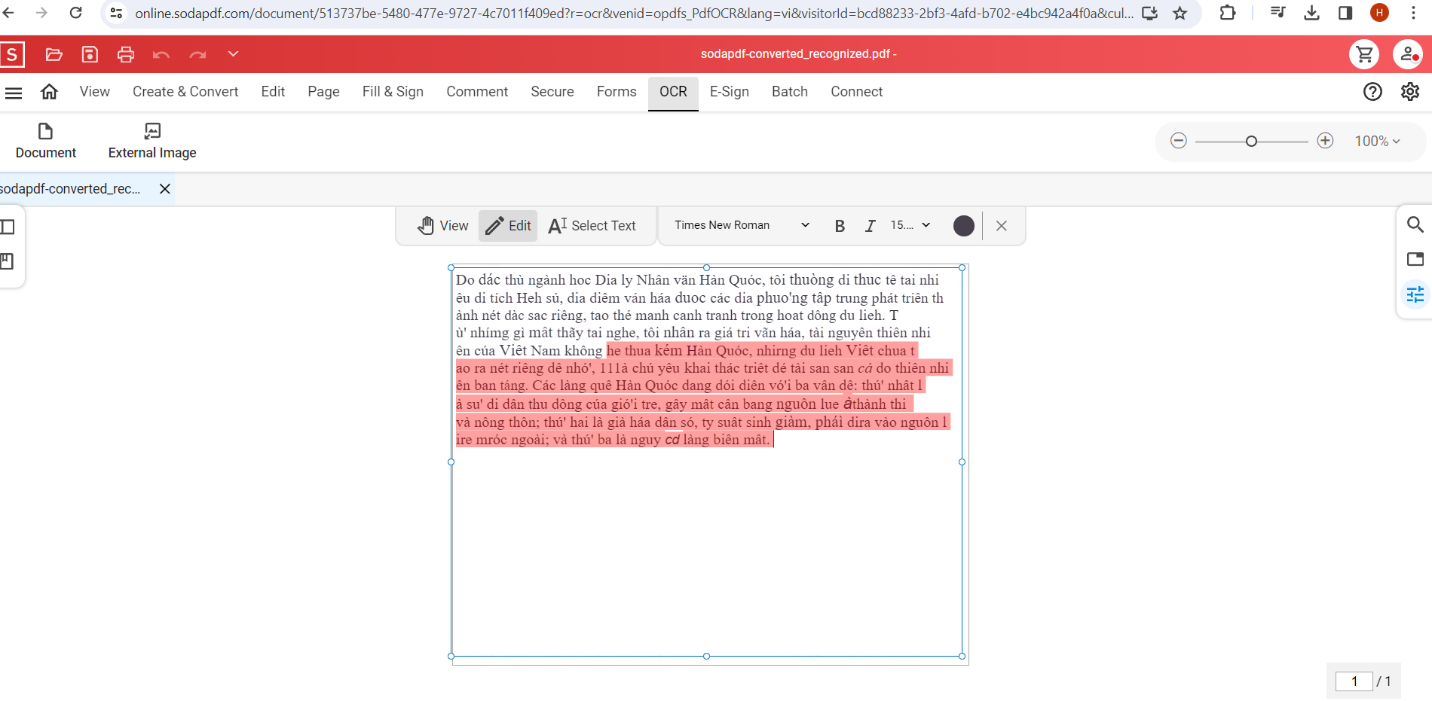
Mô tả được tạo tự động



***Chọn file hình ảnh muốn thêm vào***



***Nội dung file hình ảnh***



***Thành quả***

Nguồn các trang đã tham khảo:

<https://aws.amazon.com/vi/what-is/ocr/>

<https://nanonets.com/blog/ocr-software-best-ocr-software/>

<https://www.ibm.com/blog/optical-character-recognition/>

<https://convertio.co/vn/ocr/>

<https://www.sodapdf.com/vi/ocr-pdf/>