TRƯỜNG ĐẠI HỌC QUỐC TẾ SÀI GÒN
THE SAIGON INTERNATIONAL UNIVERSITY

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC QUỐC TẾ SÀI GÒN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN MÔN HỌC
TRÍ TUỆ NHÂN TẠO

ĐỀ TÀI: XÂY DỰNG PHẦN MỀM TÁCH TỪ VÀ HÌNH ẢNH TỪ FILE PDF

Ngành: **KHOA HỌC MÁY TÍNH**

Giảng viên hướng dẫn: **ThS. Lê Ngọc Thạch**

Sinh viên thực hiện:

- Trần Văn Đan Trường

  MSSV: 910 1180 1418 Lớp: 18DMT

  Chuyên ngành: Khoa học máy tính

Tp. Hồ Chí Minh, tháng 10 năm 2020

TRƯỜNG ĐẠI HỌC QUỐC TẾ SÀI GÒN
THE SAIGON INTERNATIONAL UNIVERSITY

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC QUỐC TẾ SÀI GÒN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN MÔN HỌC
TRÍ TUỆ NHÂN TẠO

ĐỀ TÀI: XÂY DỰNG PHẦN MỀM TÁCH TỪ VÀ HÌNH ẢNH TỪ FILE PDF

Ngành: **KHOA HỌC MÁY TÍNH**

Giảng viên hướng dẫn: **ThS. Lê Ngọc Thạch**

Sinh viên thực hiện:

- Trần Văn Đan Trường

  MSSV: 910 1180 1418 Lớp: 18DMT

  Chuyên ngành: Khoa học máy tính

Tp. Hồ Chí Minh, tháng 10 năm 2020

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

Tp. Hồ Chí Minh, ngày tháng năm 2020

Giảng viên hướng dẫn

LỜI CẢM ƠN

Để hoàn thành đồ án môn Trí tuệ nhân tạo với đề tài "Xây dựng phần mềm tách từ và hình ảnh từ file PDF", trước tiên cho phép chúng em xin gửi lời cảm ơn tới thầy Lê Ngọc Thạch đã giúp đỡ chúng em rất nhiệt tình trong suốt thời gian qua. Hơn nữa, đồ án của chúng em sẽ không thể hoàn thành tốt nếu không có sự hướng dẫn tận tình của quý thầy/cô giảng viên khoa Công nghệ thông tin - Trường Đại học Quốc tế Sài Gòn.

Thời gian thực hiện đồ án tuy ngắn, nhưng nhờ sự hướng dẫn của thầy Lê Ngọc Thạch đã tạo cơ hội cho em áp dụng nền tảng môn học Trí tuệ nhân tạo vào công tác nghiên cứu. Trong suốt thời gian hoàn thành đồ án, em đã có cơ hội rèn luyện được các kỹ năng làm việc và nâng cao hiểu biết của mình trong việc thực hiện viết báo cáo và xây dựng chương trình, từ đó nhận thức rõ hơn về tầm quan trọng của kiến thức cũng như kỹ năng thực hiện đồ án.

Vì vốn kiến thức và kinh nghiệm còn hạn chế nên bài báo cáo và chương trình thực nghiệm không thể tránh khỏi những hạn chế, thiếu sót. Chúng em rất mong muốn nhận được sự góp ý của các quý thầy/cô để giúp chúng em hoàn thiện hơn về nghiệp vụ của mình để chúng em có cơ sở, nền tảng kiến thức phục vụ cho công tác sau này với hy vọng những đồ án tiếp theo trong chương trình học được hoàn thành tốt hơn.

Một lần nữa cho phép chúng em xin chân thành cảm ơn các thầy Lê Ngọc Thạch và quý thầy/cô giảng viên khoa Công nghệ thông tin - Trường Đại học Quốc tế Sài Gòn đã tạo điều kiện cho sinh viên chúng em có cơ hội được phát triển ý tưởng sáng tạo, tìm tòi, học hỏi và biết cách áp dụng những kiến thức đã học để xây dựng ứng dụng thực tế, đó chính là kết quả của quá trình truyền đạt kiến thức của quý thầy/cô và sự trao dồi kiến thức của bản thân chúng em.

Chúng em xin cảm ơn quý thầy/cô giảng viên trong Khoa đã giúp đỡ chúng em hoàn thành đồ án và bài báo cáo này.

MỤC LỤC

DANH MỤC HÌNH ẢNH

CHƯƠNG 1. TỔNG QUAN

1.1. Thực trạng hiện nay

- Công nghệ thông tin ngày càng phát triển và có vai trò hết sức quan trọng không thể thiếu trong cuộc sống hiện đại. Con người ngày càng tạo ra những cỗ máy thông minh có khả năng tự nhận biết và xử lí được các công việc một cách tự động, phục vụ cho lợi ích của con người. Trong những năm gần đây, một trong những bài toán nhận được nhiều sự quan tâm và tốn nhiều công sức nhất của lĩnh vực công nghệ thông tin, đó chính là bài toán nhận dạng. Tuy mới xuất hiện chưa lâu nhưng nó đã rất được quan tâm vì tính ứng dụng thực tế của bài toán cũng như sự phức tạp của nó.
- PDF hay Portable Document Format là một định dạng tập tin văn bản khá phổ biến của hãng Adobe Systems. Không như văn bản Word, một văn bản PDF, trong hầu hết các trường hợp, sẽ được hiển thị giống nhau trên những môi trường làm việc khác nhau. Chính vì ưu điểm này, định dạng PDF đã trở nên phổ biến cho việc phát hành sách, báo hay các tài liệu khác qua mạng Internet. Bởi PDF không bị phụ thuộc quá nhiều như Word, PDF có thể tạo, xem bằng nhiều hãng phần mềm sản xuất khác nhau như Adobe Reader, Foxit Reader…
- Nhận dạng chữ là lĩnh vực được nhiều nhà nghiên cứu quan tâm và cho đến nay lĩnh vực này cũng đã đạt được nhiều thành tựu lớn lao cả về mặt lý thuyết lẫn ứng dụng thực tế. Lĩnh vực nhận dạng chữ được chia làm hai loại: Nhận dạng chữ in và nhận dạng chữ viết tay. Đến thời điểm này, nhận dạng chữ đã được giải quyết gần như trọn vẹn. Tuy nhiên, việc trích xuất văn bản còn gặp nhiều khó khăn vì nhiều tệp tin PDF có nội dung có kí tự thuộc mã Unicode, UTF-8, … nên sẽ nhận dạng sai. Hay một số font chữ sẽ ảnh hưởng đến kết quả của công việc trích xuất văn bản này.
- Ngoài ra, việc nhận diện hình ảnh, cũng như bảng biểu từ tệp PDF cũng là một vấn đề khó khăn do bị ảnh hưởng bởi màu sách, độ phức tạp của hình ảnh.

1.2. Nhiệm vụ đồ án

- Từ những vấn đề trên, chúng em quyết định xây dựng một phần mềm phục vụ nhu cầu trích xuất văn bản, hình ảnh,… từ file PDF nhằm góp phần thực hiện công việc chỉnh sửa, thay đổi nội dung file PDF một cách nhanh chóng, hiệu quả hơn. Hơn

nữa, sản phẩm hỗ trợ mọi đối tượng có nhu cầu sử dụng nội dung file PDF. Ngoài ra sản phẩm là tiền đề để phát triển những sản phẩm có chức năng cốt lõi là trích xuất văn bản, hình ảnh,… sau này.

1.3. Phạm vi ứng dụng

- Đơn vị sử dụng: Trường Đại học Quốc tế Sài Gòn
- Tên dự án: Xây dựng phần mềm tách từ và hình ảnh từ file PDF

1.4. Đối tượng sử dụng

- Sản phẩm được thiết kế để phục vụ cho các đối tượng: mọi đối tượng có nhu cầu sử dụng nội dung từ file PDF.

1.5. Mục tiêu của ứng dụng

- Ứng dụng được thiết kế để thực hiện:
  + Không phải gõ nội dung văn bản từ file PDF một cách thủ công.
  + Một file PDF có nội dung lớn thì công việc gõ thủ công là việc không thể. Hơn nữa độ chính xác chưa được đảm bảo.
  + Ứng dụng tin học hóa vào công việc.
  + Tiết kiệm được thời gian so với việc gõ nội dung thủ công thủ công.

1.6. Các bước xây dựng đồ án

- Lập kế hoạch phát triển hệ thống
- Phân tích hệ thống.
- Thiết kế
- Cài đặt
- Kiểm tra
- Biên soạn tài liệu và hướng dẫn

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1. C Sharp (C#)

2.1.1. Tổng quan về C#

2.1.1.1. C# là gì?

- C# là một ngôn ngữ lập trình đơn giản, được phát triển bởi đội ngũ kỹ sư của Microsoft vào năm 2000, trong đó người dẫn đầu là Anders Hejlsberg và Scott Wiltamuth.
- C# là ngôn ngữ lập trình hiện đại, hướng đối tượng và nó được xây dựng trên nền tảng của hai ngôn ngữ mạnh nhất là C++ và Java.
- C# được thiết kế cho Common Language Infrastructure (CLI), mà gồm Executable Code và Runtime Environment, cho phép chúng ta sử dụng các ngôn ngữ high-level đa dạng trên các nền tảng và cấu trúc máy tính khác nhau.
- C# với sự hỗ trợ mạnh mẽ của .NET Framework giúp cho việc tạo một ứng dụng Windows Forms hay WPF (Windows Presentation Foundation), . . . trở nên rất dễ dàng.

2.1.1.2. Đặc trưng của C#

- C# là ngôn ngữ đơn giản: như ta đã biết thì ngôn ngữ C# dựng trên nền tảng C++ và Java nên ngôn ngữ C# khá đơn giản. Nếu chúng ta thân thiện với C và C++ hoặc thậm chí là Java, chúng ta sẽ thấy C# khá giống về diện mạo, cú pháp, biểu thức, toán tử và những chức năng khác được lấy trực tiếp từ ngôn ngữ C và C++, nhưng nó đã được cải tiến để làm cho ngôn ngữ đơn giản hơn. Một vài trong các sự cải tiến là loại bỏ các dư thừa, hay là thêm vào những cú pháp thay đổi.
- C# là ngôn ngữ hiện đại : xử lý ngoại lệ, những kiểu dữ liệu mở rộng, bảo mật mã nguồn..v..v....
- C# là một ngôn ngữ lập trình thuần hướng đối tượng.
- C# là một ngôn ngữ ít từ khóa
- Ưu điểm nổi bật của C#:
  + C# có cấu trúc khá gần gũi với các ngôn ngữ lập trình truyền thống, nên cũng khá dể dàng tiếp cận và học nhanh với C#.
  + C# có thể biên dịch trên nhiều nền tảng máy tính khác nhau.

+ C# được xây dựng trên nền tảng của C++ và Java nên nó được thừa hưởng những ưu điểm của ngôn ngữ đó.
+ C# là một phần của .NET Framework nên được sự chống lưng khá lớn đến từ bộ phận này.
+ C# có IDE Visual Studio cùng nhiều plug-in vô cùng mạnh mẽ.

2.1.2. *Môi trường lập trình C#*

- Microsoft Visual Studio là một môi trường phát triển tích hợp (IDE) từ Microsoft. Nó được sử dụng để phát triển chương trình máy tính cho Microsoft Windows, cũng như các trang web, các ứng dụng web và các dịch vụ web. Visual Studio sử dụng nền tảng phát triển phần mềm của Microsoft như Windows API, Windows Forms, Windows Presentation Foundation, Windows Store và Microsoft Silverlight. Nó có thể sản xuất cả hai ngôn ngữ máy và mã số quản lý.
- Visual Studio bao gồm một trình soạn thảo mã hỗ trợ IntelliSense cũng như cải tiến mã nguồn. Trình gỡ lỗi tích hợp hoạt động cả về trình gỡ lỗi mức độ mã nguồn và gỡ lỗi mức độ máy. Công cụ tích hợp khác bao gồm một mẫu thiết kế các hình thức xây dựng giao diện ứng dụng, thiết kế web, thiết kế lớp và thiết kế giản đồ cơ sở dữ liệu. Nó chấp nhận các plug-in nâng cao các chức năng ở hầu hết các cấp bao gồm thêm hỗ trợ cho các hệ thống quản lý phiên bản (như Subversion) và bổ sung thêm bộ công cụ mới như biên tập và thiết kế trực quan cho các miền ngôn ngữ cụ thể hoặc bộ công cụ dành cho các khía cạnh khác trong quy trình phát triển phần mềm.
- Visual Studio hỗ trợ nhiều ngôn ngữ lập trình khác nhau và cho phép trình biên tập mã và gỡ lỗi để hỗ trợ (mức độ khác nhau) hầu như mọi ngôn ngữ lập trình. Các ngôn ngữ tích hợp gồm có C, C++ và C++/CLI (thông qua Visual C++), VB.NET (thông qua Visual Basic.NET), C# (thông qua Visual C#) và F# (như của Visual Studio 2010). Hỗ trợ cho các ngôn ngữ khác như J++/J#, Python và Ruby thông qua dịch vụ cài đặt riêng rẽ. Nó cũng hỗ trợ XML/XSLT, HTML/XHTML, JavaScript và CSS.

2.2. Thư viện iTextSharp

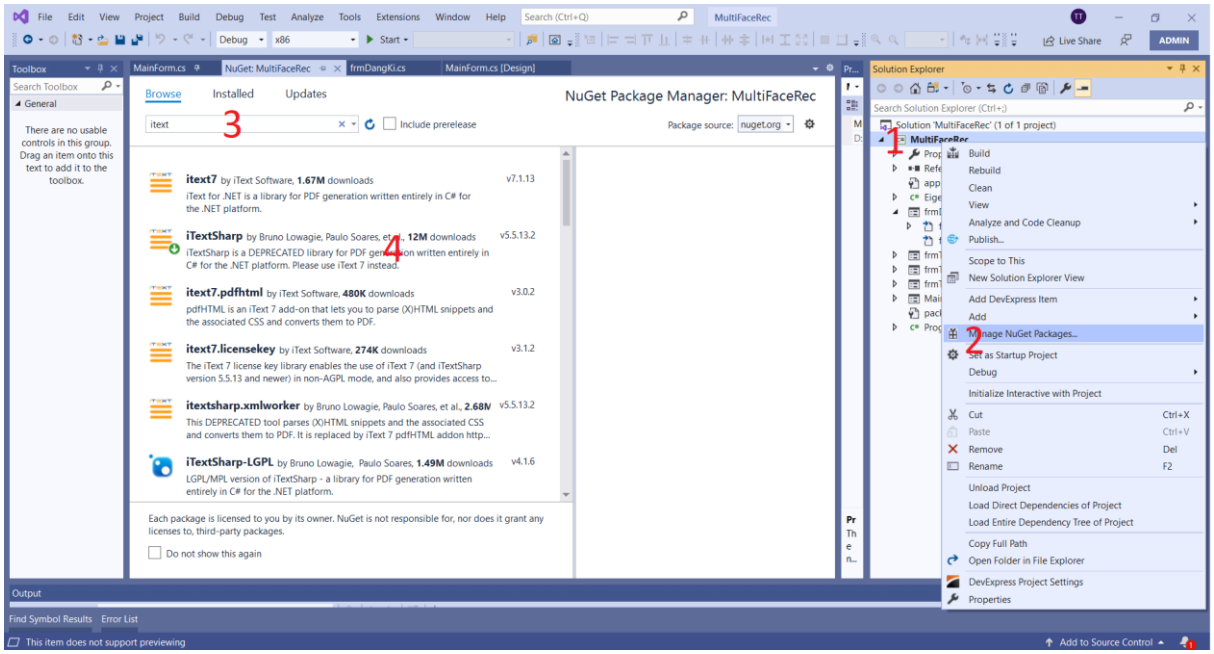
2.2.1. *iText là gì?*

- iText là một thư viện cho phép bạn tạo và thao tác các tài liệu PDF. Nó cho phép các lập trình viên tìm cách để mở rộng tính năng của web và các ứng dụng khác.
- iText dùng trong Java và .NET.
- iText được viết bởi Bruno Lowagie.
- iText cung cấp hỗ trợ cho hầu hết các tính năng PDF nâng cao như chữ ký dựa trên PKI, mã hóa 40 bit và 128 bit, sửa màu, PDF được gắn thẻ, biểu mẫu PDF (AcroForms), PDF/X, quản lý màu thông qua hồ sơ ICC và mã vạch, và được sử dụng bởi một số sản phẩm và dịch vụ, bao gồm Eclipse BIRT, Jasper Reports, JBoss Seam, Windward Reports và pdftk.

2.2.2. *Các tiện ích của iTextSharp*

- Hỗ trợ file PDF trong trình duyệt
- Tạo ra các tài liệu động từ các tập tin XML hoặc cơ sở dữ liệu
- Sử dụng nhiều tính năng tương tác PDF
- Thêm dấu trang, số trang, hình mờ,...
- Chia, nối, và thao tác các trang PDF
- Tự động điền vào các mẫu đơn PDF
- Thêm chữ ký số vào một tập tin PDF
- iText có sẵn trong Java cũng như trong C #.

2.2.3. *Cách cài đặt iTextSharp*



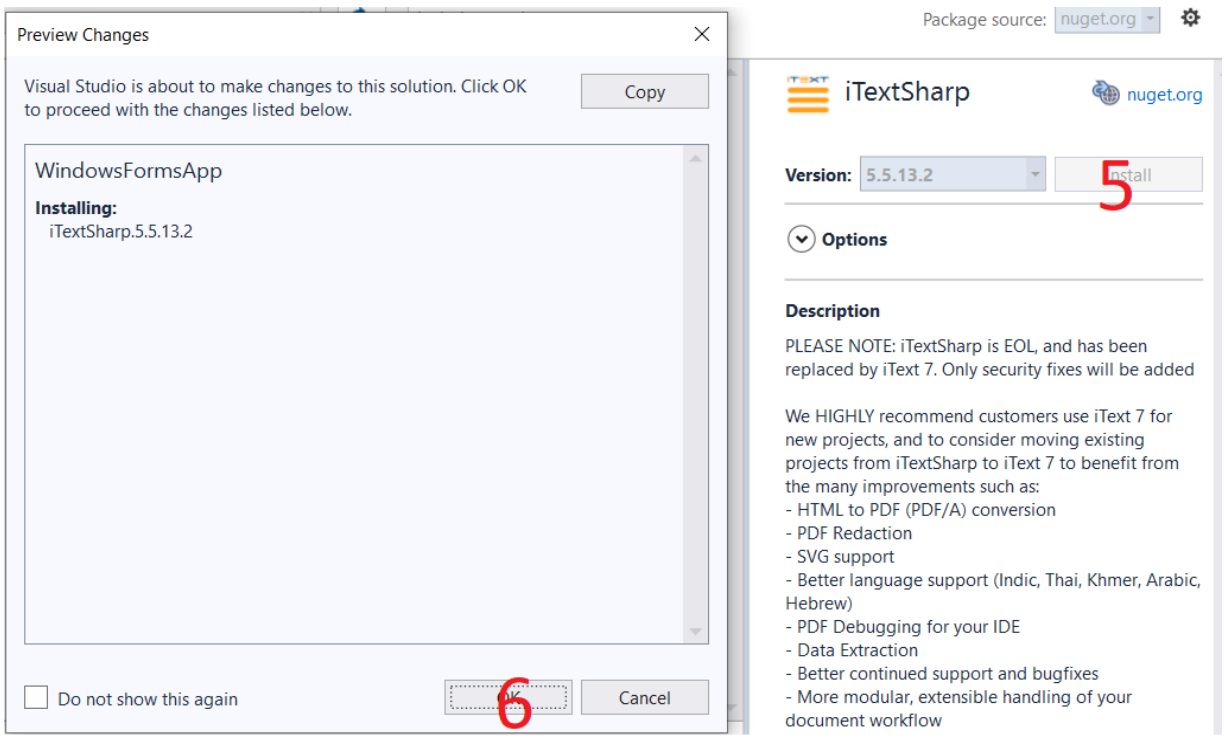
Hình 2.1: Thao tác tìm kiếm iTextSharp

– Bước 1: Chuột phải vào project hiện hành

– Bước 2: Chọn “Manage NuGet Packages…”

– Bước 3: Ở thẻ “Browse” gõ vào “iTextSharp”

– Bước 4: Chọn thư viện “iTextSharp”



Hình 2.2: Cài đặt thư viện iTextSharp

– Bước 5: Chọn “Install”
– Bước 6: Xuất hiện hộp thoại chọn “OK” -> “I Accept”.
– Bước 7: Thư viện đã được cài đặt thành công vào project.

2.3. Thư viện Spire

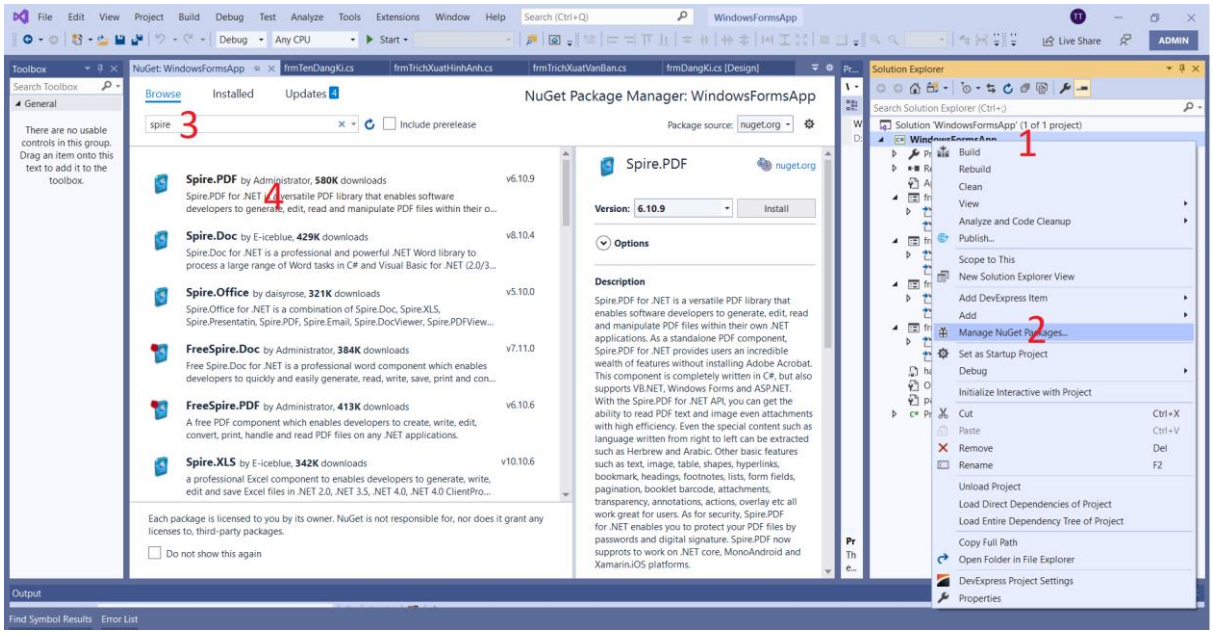
2.3.1. *Spire là gì?*

– Spire.PDF for .NET là một API PDF chuyên nghiệp được áp dụng để tạo, viết, chỉnh sửa, xử lý và đọc các tệp PDF mà không có bất kỳ phụ thuộc bên ngoài nào trong ứng dụng .NET (C #, VB.NET, ASP.NET, .NET Core, Xamarin).
– Sử dụng thư viện PDF .NET này, bạn có thể triển khai các khả năng phong phú để tạo tệp PDF từ đầu hoặc xử lý các tài liệu PDF hiện có hoàn toàn thông qua C # / VB.NET mà không cần cài đặt Adobe Acrobat.

2.3.2. *Ưu điểm của Spire*

– Nhiều tính năng phong phú có thể được hỗ trợ bởi .NET PDF API, chẳng hạn như thêm chữ ký điện tử, bao gồm dấu thời gian trong chữ ký, tạo Danh mục đầu tư PDF, trích xuất văn bản / tệp đính kèm / hình ảnh PDF, hợp nhất / tách PDF, cập nhật siêu dữ liệu, phần, vẽ đồ thị / hình ảnh và chèn, tạo và xử lý bảng và nhập dữ liệu, v.v.
– Ngoài ra, Spire.PDF for .NET có thể được áp dụng để dễ dàng chuyển đổi Văn bản, Hình ảnh, SVG, HTML sang PDF và chuyển đổi PDF sang Excel với C # / VB.NET với chất lượng cao.

2.3.3. *Cài đặt thư viện Spire*



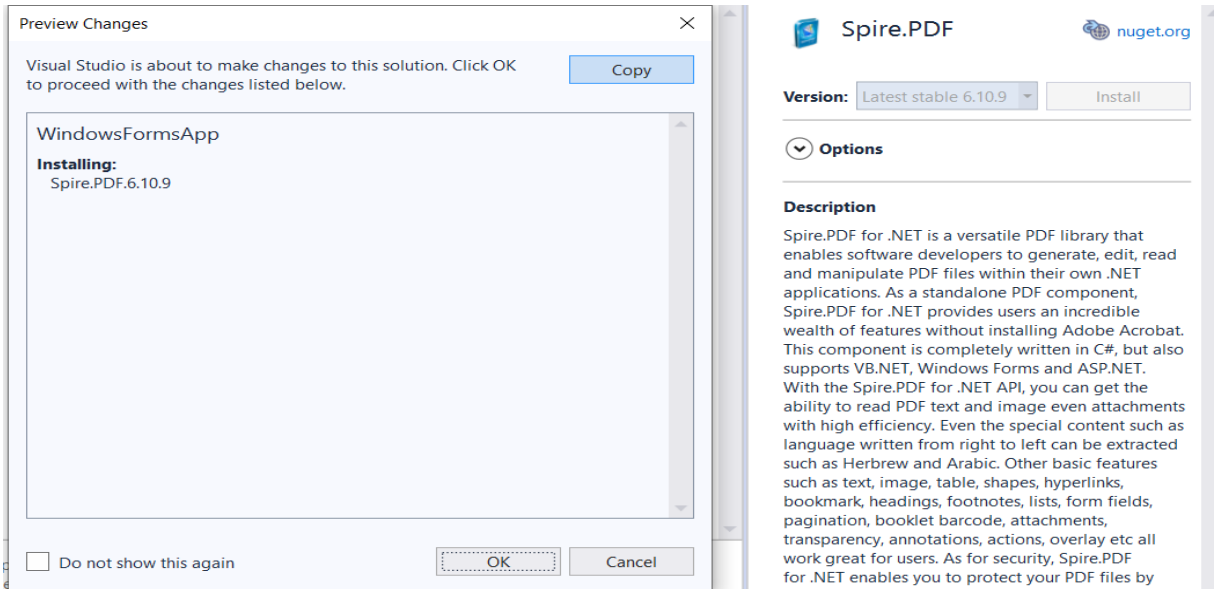
Hình 2.3: Thao tác tìm kiếm thư viện Spire

– Bước 1: Chuột phải vào project hiện hành

– Bước 2: Chọn “Manage NuGet Packages…”

– Bước 3: Ở thẻ “Browse” gõ vào “Spire”

– Bước 4: Chọn thư viện “Spire.PDF”



Hình 2.4: Cài đặt thư viện Spire

– Bước 5: Chọn “Install”

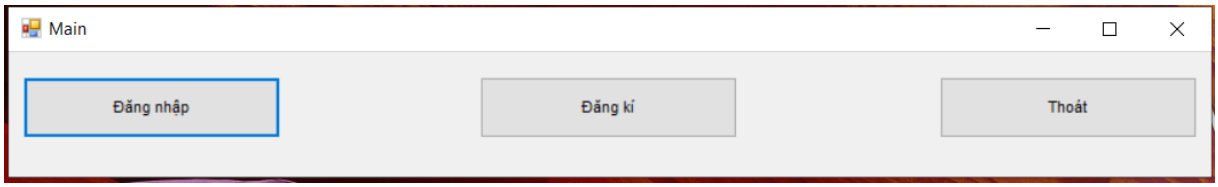
– Bước 6: Xuất hiện hộp thoại chọn “OK” -> “I Accept”.

– Bước 7: Thư viện đã được cài đặt thành công vào project.

CHƯƠNG 3. KẾT QUẢ THỰC NGHIỆM

3.1. Các thiết kế hệ thống

3.1.1. *Biểu mẫu Khởi động chương trình*

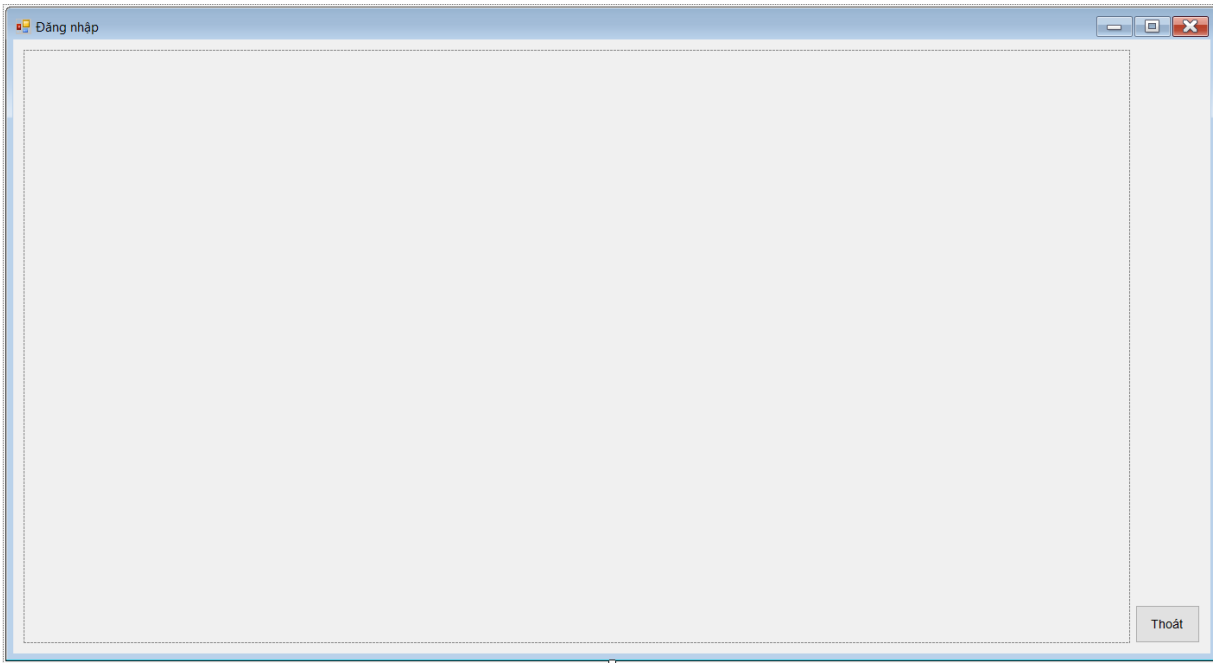


Hình 3.1: Biểu mẫu Khởi động chương trình

Bảng 3.1: Bảng chi tiết biểu mẫu Khởi động chương trình

STT	Tên Control	Loại	Ý nghĩa
1	Btndangnhap	Button	Đăng nhập vào chương trình
2	Btndangki	Button	Đăng kí tài khoản
3	Btnthoat	Button	Thoát chương trình

3.1.2. *Biểu mẫu Đăng nhập*

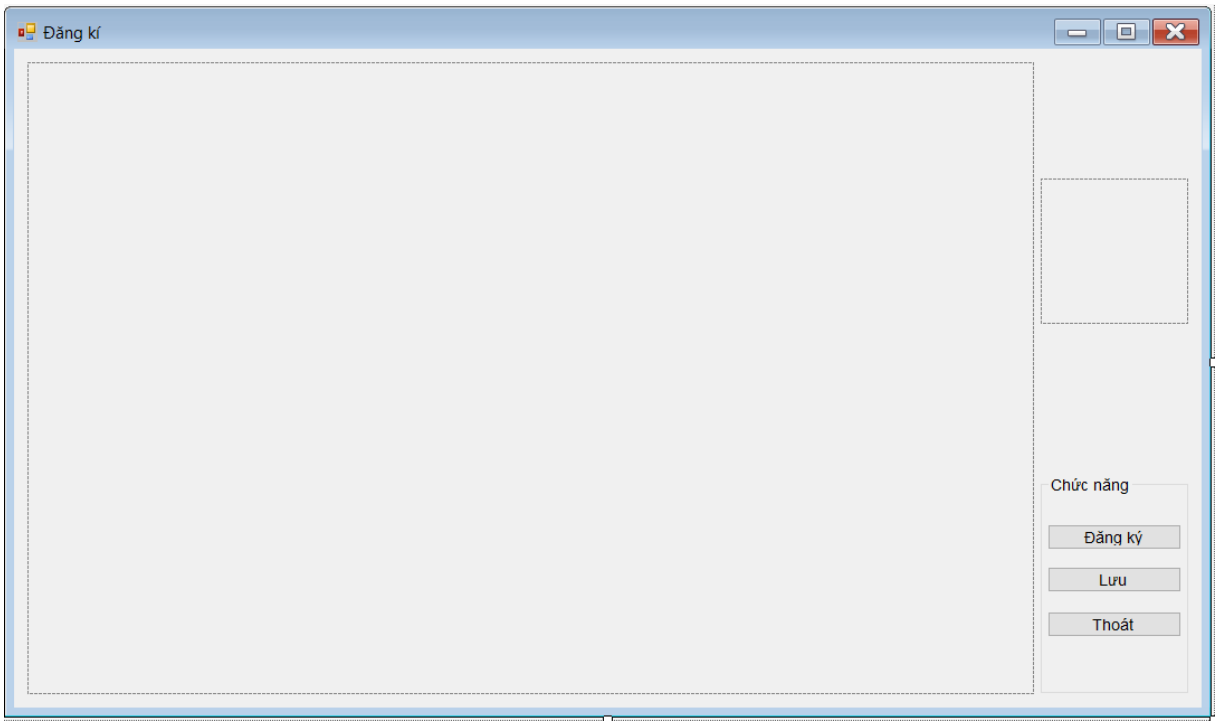


Hình 3.2: Biểu mẫu Đăng nhập

Bảng 3.2: Bảng chi tiết biểu mẫu Đăng nhập

STT	Tên Control	Loại	Ý nghĩa
1	btnThoat	Button	Thoát chương trình
2	picFace	PictureBox	Hiển thị hình ảnh quan sát được qua camera

3.1.3. *Biểu mẫu Đăng kí*

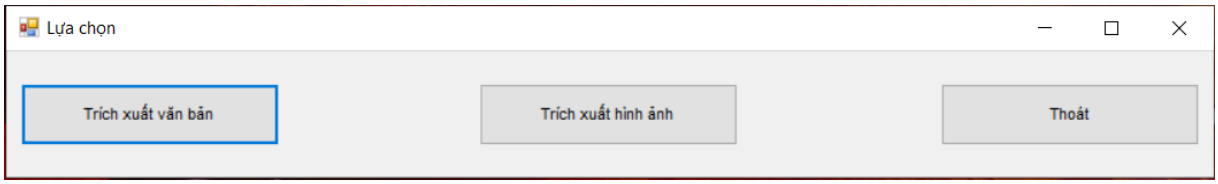


Hình 3.3: Biểu mẫu Đăng kí

Bảng 3.3: Bảng chi tiết biểu mẫu Đăng kí

STT	Tên Control	Loại	Ý nghĩa
1	picFace	PictureBox	Hiển thị hình ảnh quan sát được qua camera
2	picDetected	PictureBox	Hiển thị khuôn mặt đã đăng kí
3	btnDangKi	Button	Đăng kí khuôn mặt
4	btnLuu	Button	Lưu khuôn mặt đã đăng kí
5	btnThoat	Button	Thoát chương trình

3.1.4. *Biểu mẫu Lựa chọn chức năng*

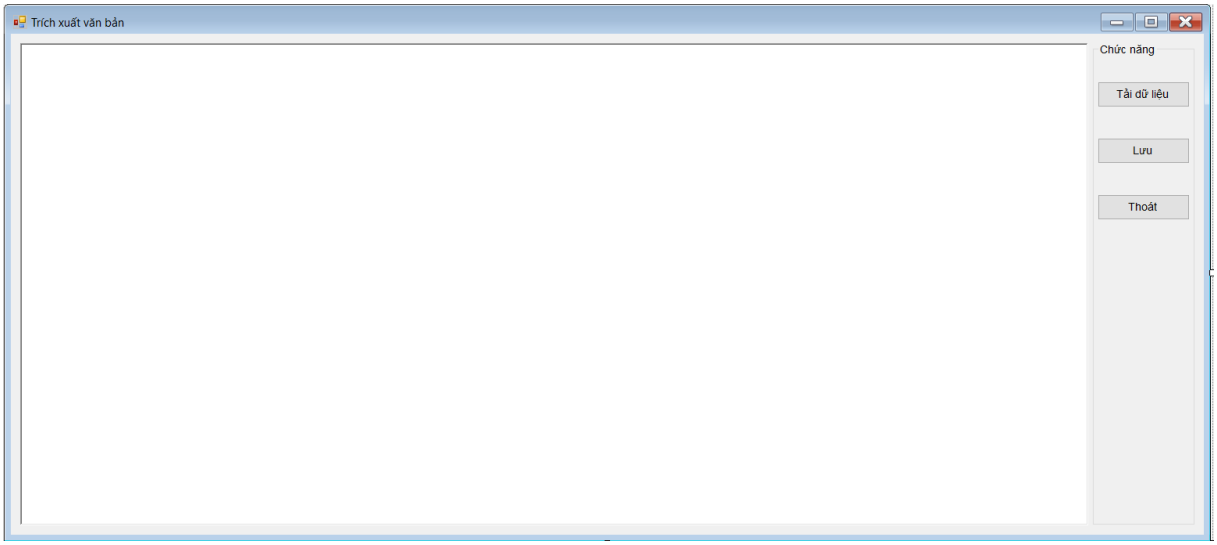


Hình 3.4: Biểu mẫu Lựa chọn chức năng

Bảng 3.4: Bảng chi tiết biểu mẫu Lựa chọn chức năng

STT	Tên Control	Loại	Ý nghĩa
1	btnVanban	Button	Mở chương trình trích xuất văn bản
2	btnHinhanh	Button	Mở chương trình trích xuất hình ảnh
3	btnThoat	Button	Thoát chương trình

3.1.5. *Biểu mẫu Trích xuất văn bản*

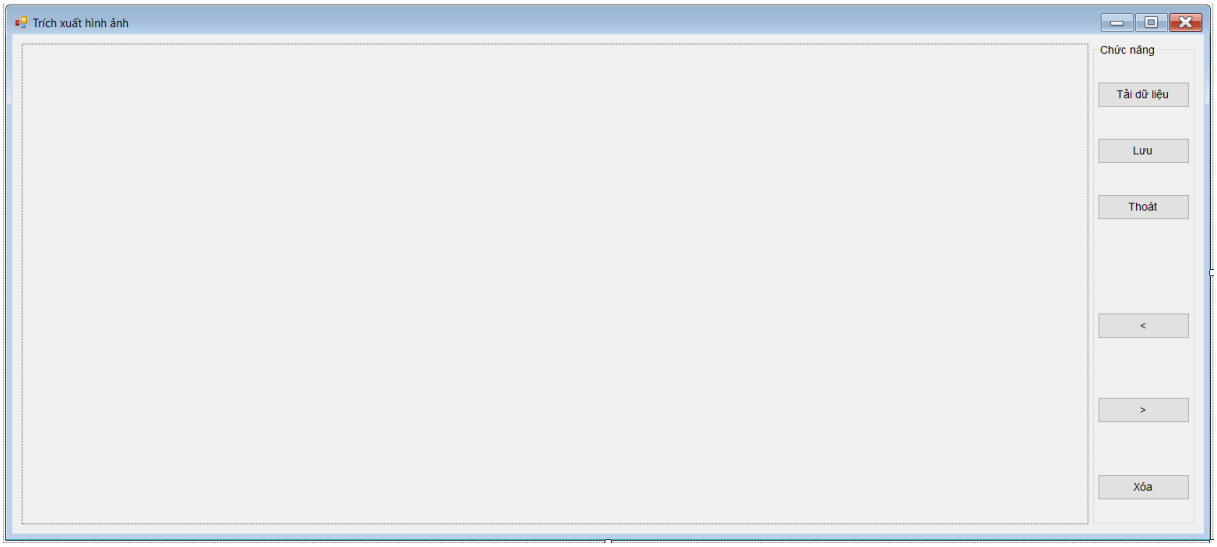


Hình 3.5: Biểu mẫu Trích xuất văn bản

Bảng 3.5: Bảng chi tiết biểu mẫu Trích xuất văn bản

STT	Tên Control	Loại	Ý nghĩa
1	txtVanban	Textbox	Văn bản được trích xuất
2	btnTaidulieu	Button	Chọn file PDF để trích xuất
3	btnLuu	Button	Lưu văn bản đã được trích xuất
4	btnThoat	Button	Thoát chương trình

3.1.6. *Biểu mẫu Trích xuất hình ảnh*



Hình 3.6: Biểu mẫu Trích xuất hình ảnh

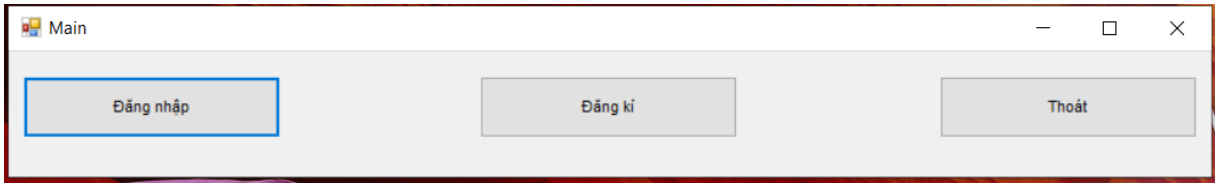
Bảng 3.6: Bảng chi tiết biểu mẫu Trích xuất hình ảnh

STT	Tên Control	Loại	Ý nghĩa
1	picHinhanh	PictureBox	Hình ảnh được trích xuất

2	btnTaidulieu	Button	Chọn file PDF để trích xuất
3	btnLuu	Button	Lưu hình ảnh đã được trích xuất
4	btnThoat	Button	Thoát chương trình
5	btnForward	Button	Hiển thị hình ảnh kế tiếp
6	btnBackward	Button	Hiển thị hình ảnh trước đó
7	btnXoa	Button	Xóa hình ảnh hiện hành

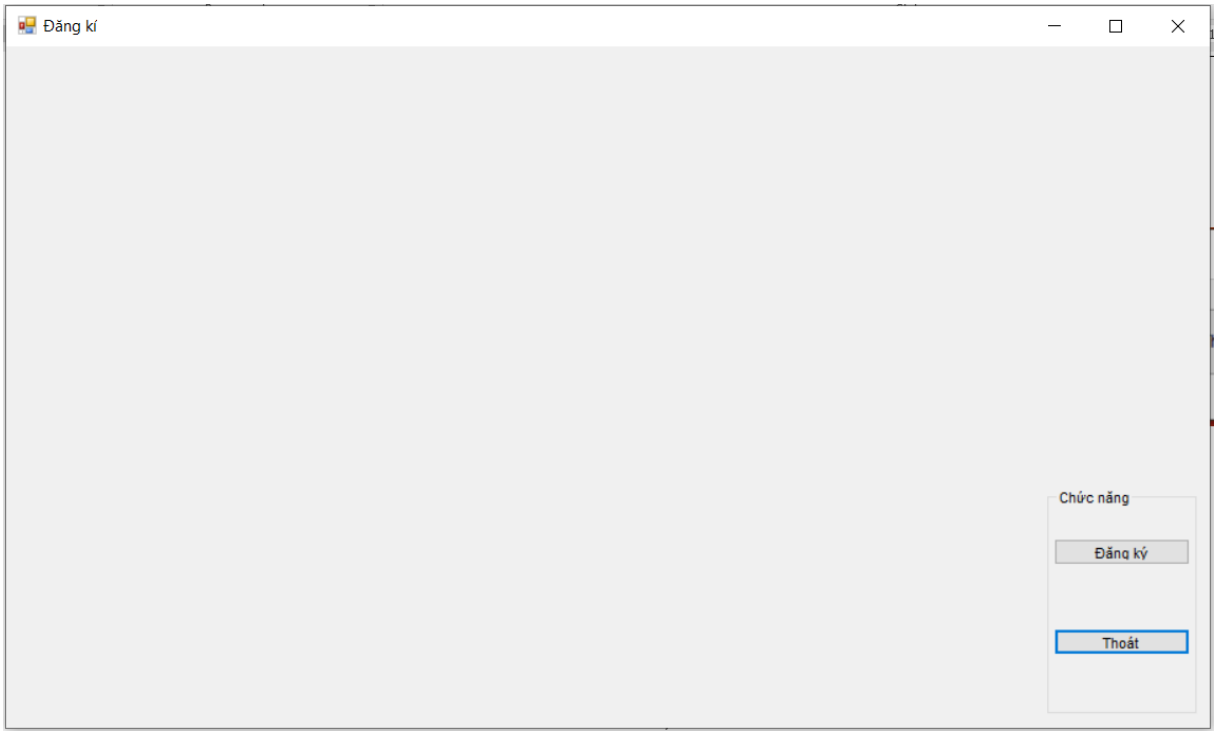
3.2. Quy trình xử dụng

- Khi khởi động chương trình sẽ có giao diện như bên dưới



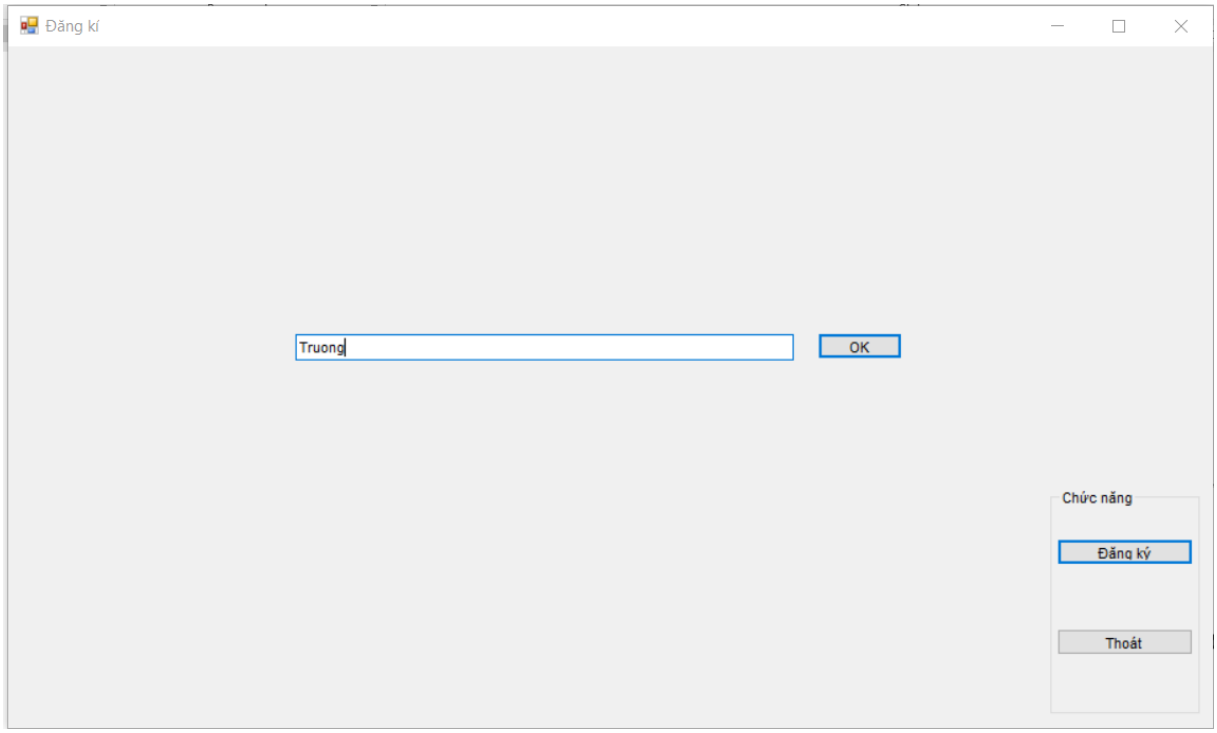
Hình 3.7: Giao diện khi khởi động chương trình

- Ở đây sẽ có 2 lựa chọn cho người dùng:
  - Đối với người dùng đã có tài khoản sẽ lựa chọn "Đăng nhập"
  - Đối với người dùng mới sẽ lựa chọn "Đăng kí"
- Khi lựa chọn "Đăng kí", chương trình sẽ xuất hiện giao diện cho người dùng đăng kí



Hình 3.8: Giao diện đăng kí

- Người dùng chọn "Đăng ký", sẽ xuất hiện hộp thoại để nhập tên tài khoản và nhấn "OK" hay phím "Enter" để kết thúc quá trình đăng kí tên tài khoản.

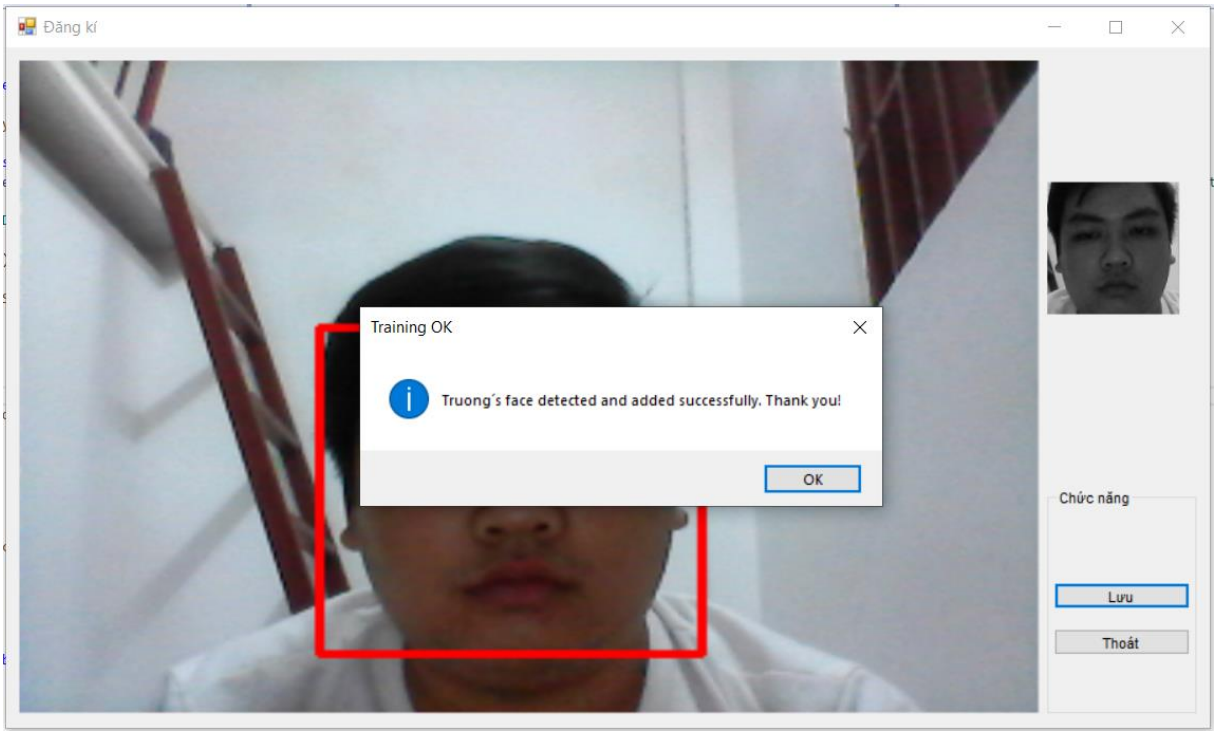


Hình 3.9: Đăng kí tên tài khoản

- Sau khi đã đăng kí tên tài khoản, chương trình sẽ mở camera để ghi nhận khuôn mặt

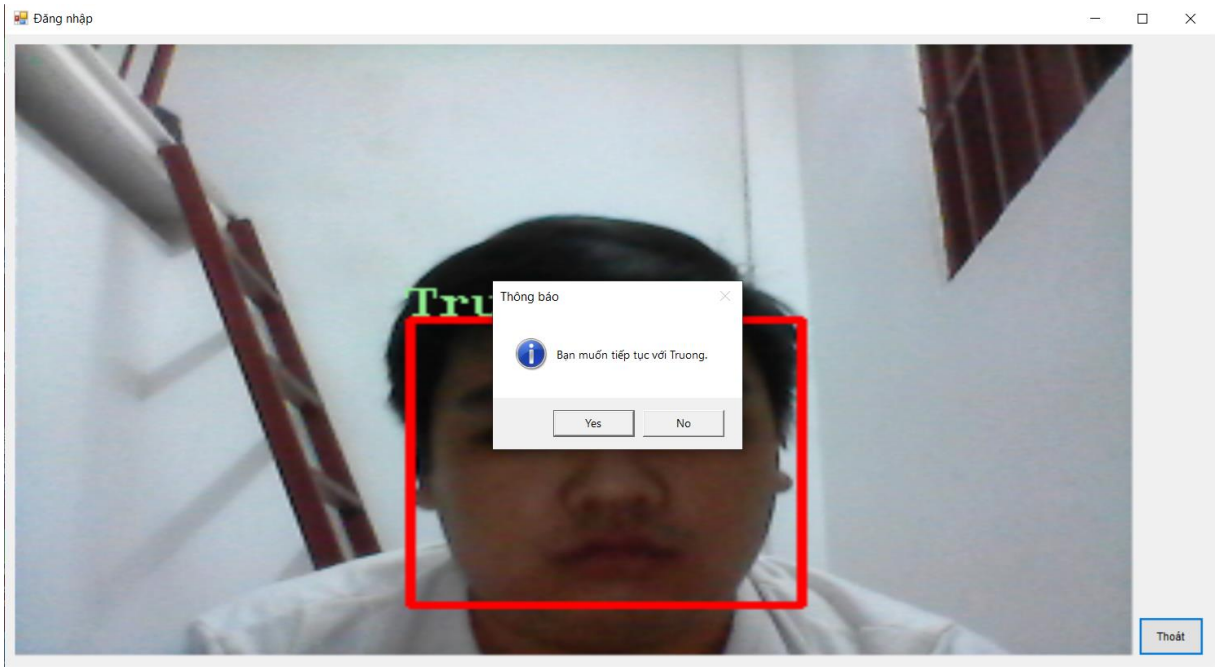
Hình 3.10: Phát hiện khuôn mặt

- Sau khi phát hiện được khuôn mặt, người dùng chọn “Lưu”, khi này chương trình sẽ báo lưu tài khoản thành công



Hình 3.11: Thông báo đã tạo tài khoản thành công

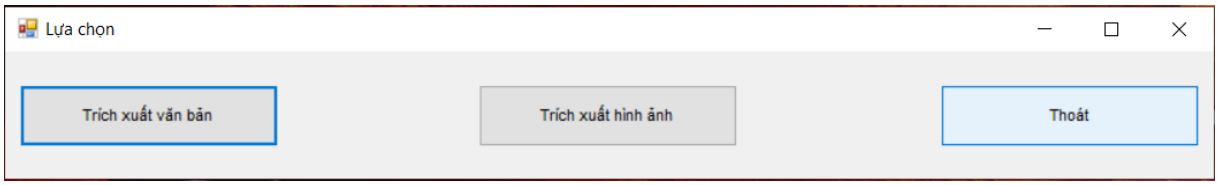
- Bây giờ chương trình sẽ tự động đăng nhập cho người dùng, ở bước này tương đương với việc người dùng chọn "Đăng nhập" khi khởi động chương trình. Chương trình sẽ nhận khuôn mặt và xác nhận tên người dùng



Hình 3.12: Thông báo đăng nhập thành công

- Tiếp theo chương trình sẽ xuất hiện các công việc cho người dùng như hình bên dưới:
  + Trích xuất văn bản từ file PDF

+ Trích xuất hình ảnh từ file PDF



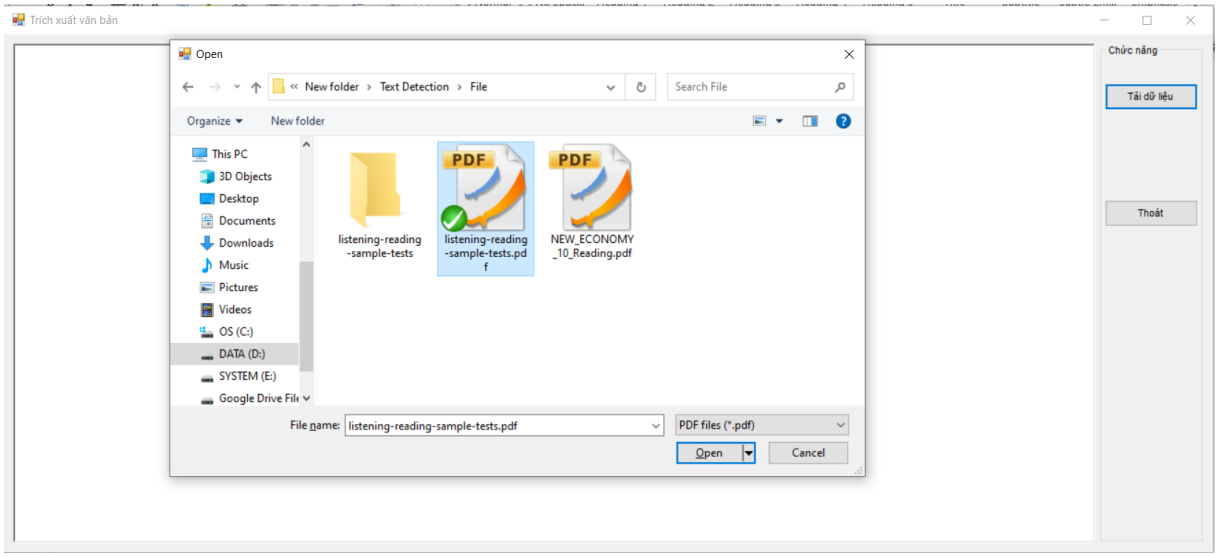
Hình 3.13: Giao diện làm việc

❖ Đối với công việc Trích xuất văn bản

- Khi người dùng chọn "Trích xuất văn bản", chương trình sẽ xuất hiện giao diện làm việc trích xuất văn bản từ file PDF

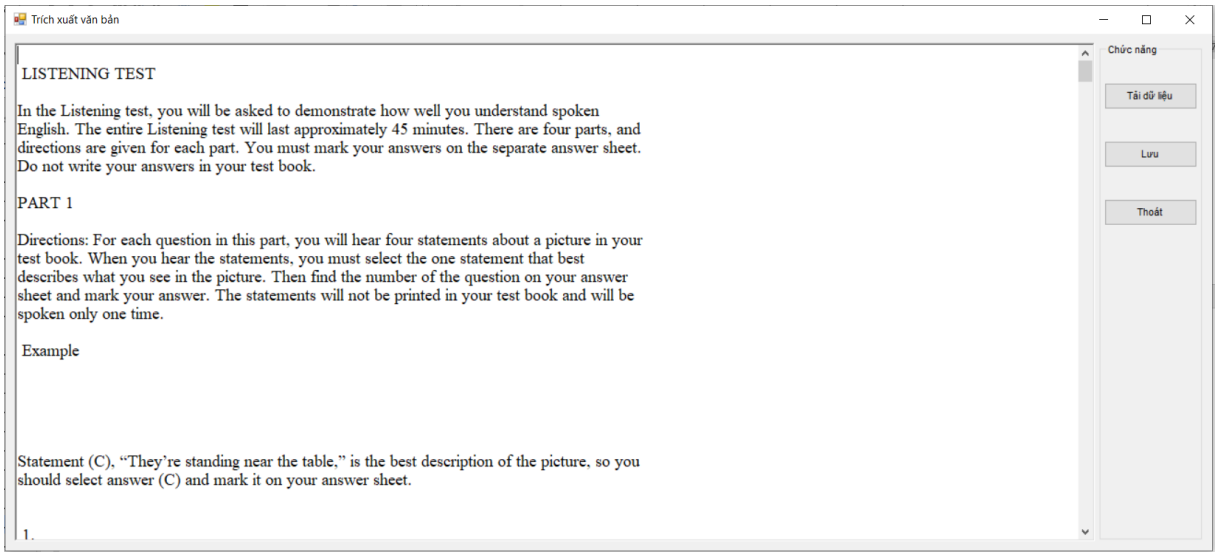
Hình 3.14: Giao diện công việc trích xuất văn bản

- Người dùng sẽ chọn tiếp "Tải dữ liệu", chương trình sẽ hiện ra hộp thoại để người dùng chọn file PDF mà có nhu cầu trích xuất, và nhấn "Open" để chọn



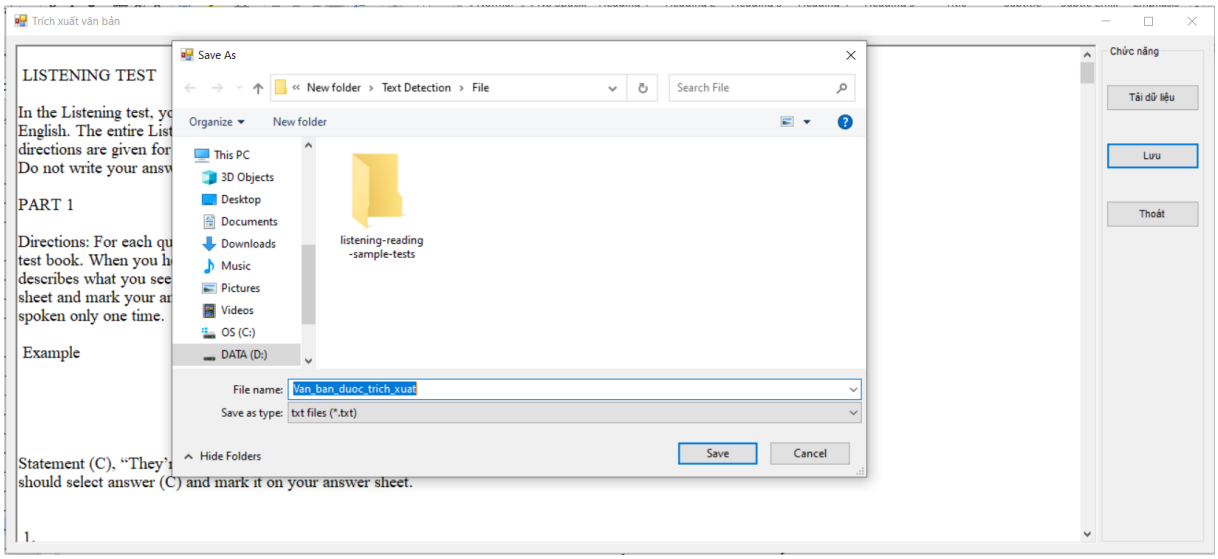
Hình 3.15: Chọn file PDF

- Sau khi chọn được file, chương trình sẽ trích xuất văn bản từ file PDF đã chọn, kết quả trả về sẽ được hiện trong TextBox như hình bên dưới



Hình 3.16: Kết quả việc trích xuất văn bản

- Người dùng có thể chọn "Lưu" để lưu văn bản được trích xuất, khi nhấn "Lưu", chương trình sẽ hiện ra hộp thoại để người dùng chọn địa chỉ để lưu văn bản. Văn bản được lưu dưới đuôi .txt, nhấn "Save" để hoàn thành thao tác.

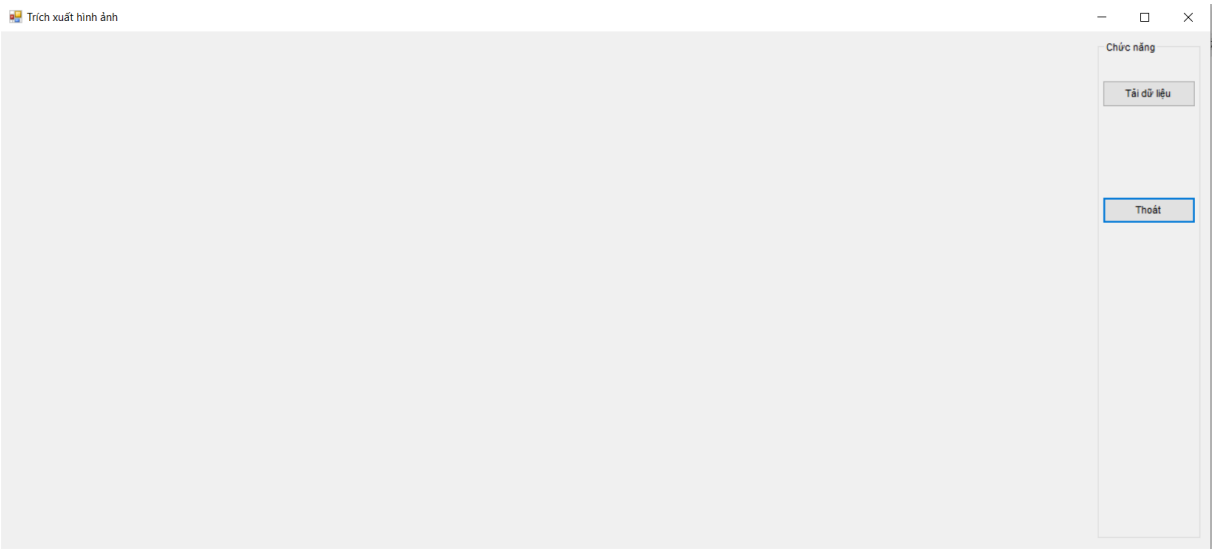


Hình 3.17: Lưu văn bản

- Đến đây thao tác trích xuất văn bản từ file PDF đã hoàn thành.

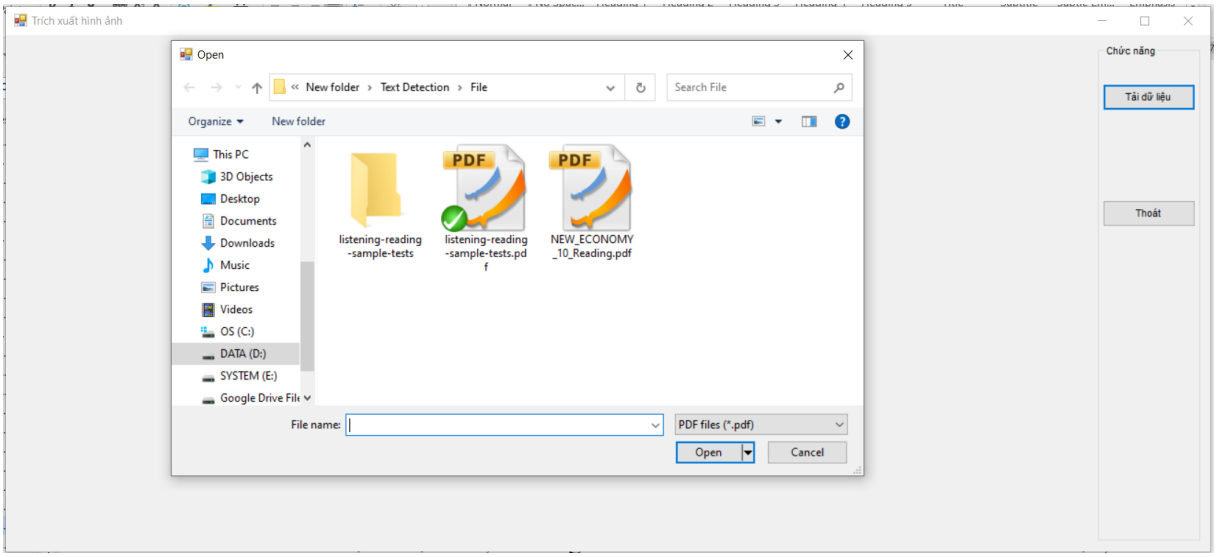
❖ Đối với công việc Trích xuất hình ảnh

- Khi người dùng chọn "Trích xuất hình ảnh", chương trình sẽ xuất hiện giao diện làm việc trích xuất hình ảnh từ file PDF



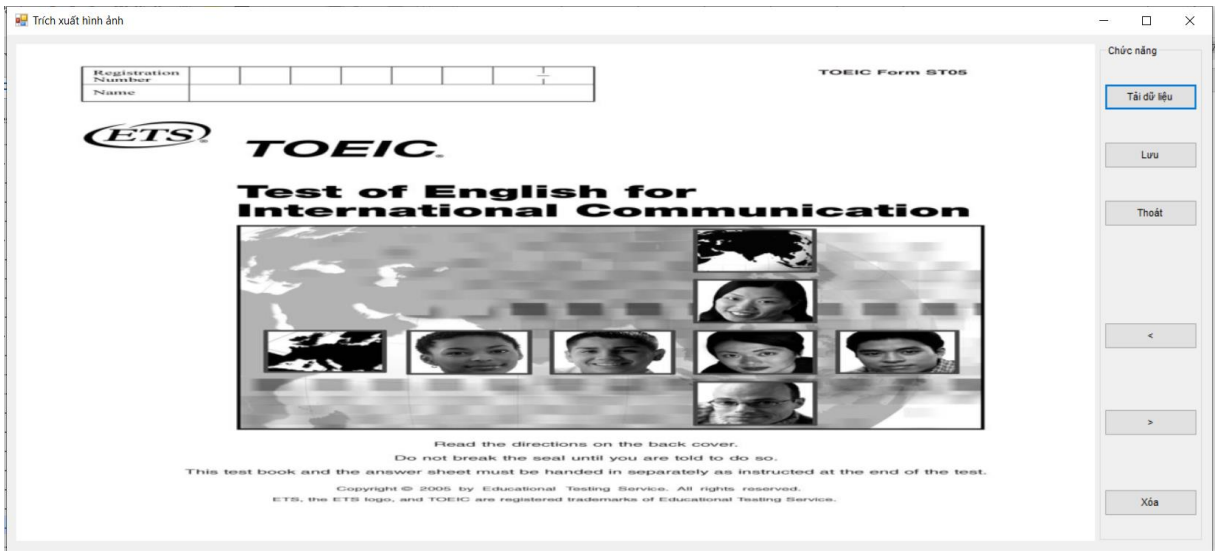
Hình 3.18: Giao diện công việc trích xuất hình ảnh

- Người dùng sẽ chọn tiếp "Tải dữ liệu", chương trình sẽ hiện ra hộp thoại để người dùng chọn file PDF mà có nhu cầu trích xuất, và nhấn "Open" để chọn



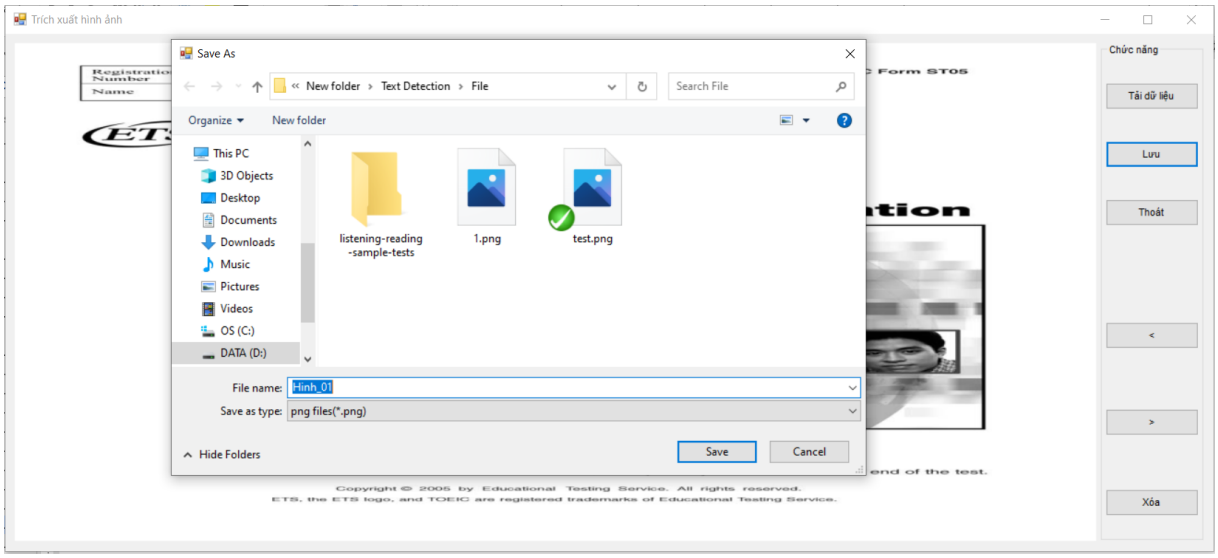
Hình 3.19: Chọn file PDF

- Sau khi chọn được file, chương trình sẽ trích xuất hình ảnh từ file PDF đã chọn, kết quả trả về sẽ được hiện trong PictureBox như hình bên dưới



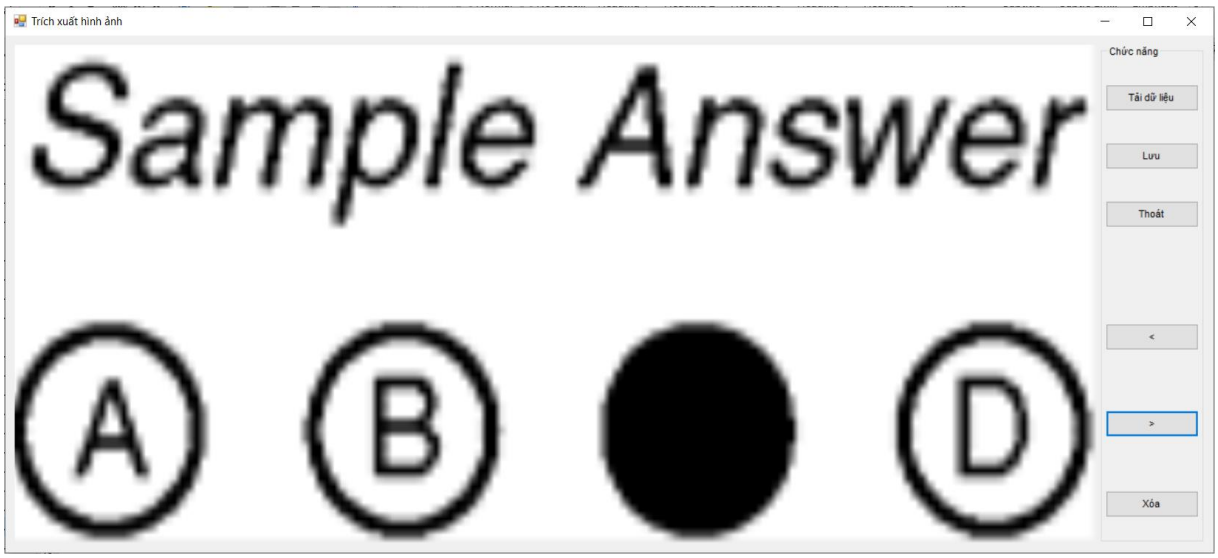
Hình 3.20: Kết quả việc trích xuất hình ảnh

- Người dùng có thể chọn "Lưu" để lưu hiện hành ở PictureBox, khi nhấn "Lưu", chương trình sẽ hiện ra hộp thoại để người dùng chọn địa chỉ để lưu hình ảnh. Hình ảnh được lưu dưới đuôi .png, nhấn "Save" để hoàn thành thao tác.



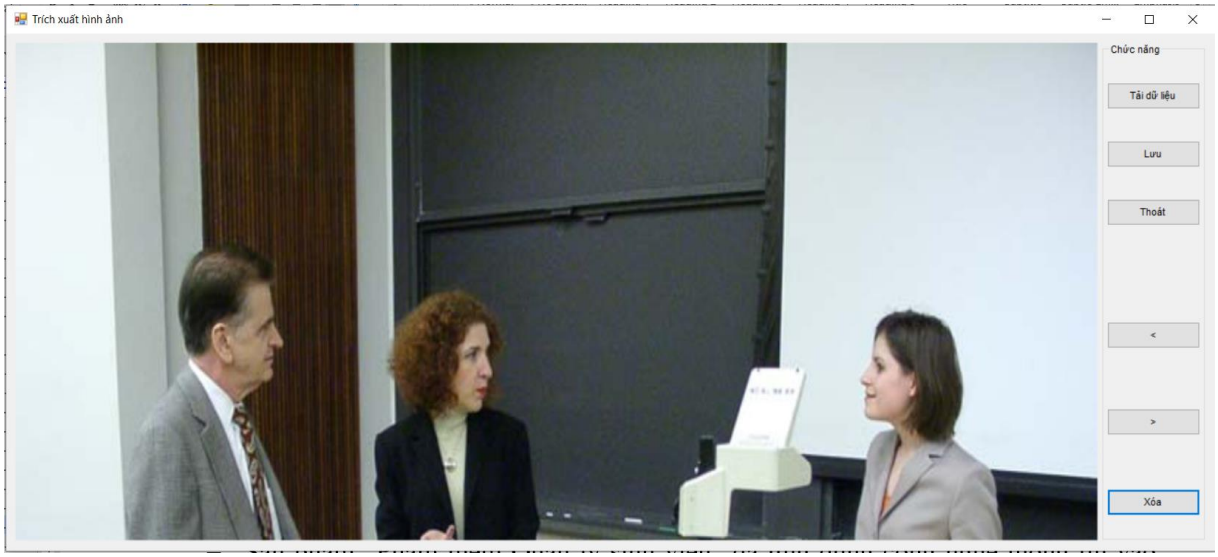
Hình 3.21: Lưu hình ảnh

- Người dùng có thể chọn ">" hoặc "<" để di chuyển hình ảnh hiện hành "tiến" hoặc "lùi" một hình, hình ảnh hiện hành sẽ được cập nhật ở PictureBox



Hình 3.22: "Tiến" hình ảnh

- Người dùng có thể xóa hình ảnh hiện hành ở PictureBox bằng cách chọn "Xóa", tuy nhiên, hình ảnh đã xóa không thể phục hồi. Danh sách hình ảnh sẽ được cập nhật sau khi xóa.



Hình 3.23: Thao tác xóa hình ảnh hiện hành

- Đến đây thao tác trích xuất hình ảnh từ file PDF đã hoàn thành.

CHƯƠNG 4. KẾT LUẬN

4.1. Kết quả đạt được

4.1.1. *Ưu điểm*

- Sản phẩm "Phần mềm tách từ và hình ảnh từ file PDF" đã ứng dụng công nghệ thông tin vào công tác xử lý văn bản, giúp nâng cao hiệu quả và năng suất làm việc so với công tác tổ chức truyền thống.
- Các công việc xóa, sửa, lưu trữ,… được tổ chức chặt chẽ, chính xác.
- Sử dụng dễ dàng, thành thạo nhanh chóng, tốn ít thời gian.
- Chỉ cần một người sử dụng sản phẩm là có thể hoàn thành công việc được ngay.
- Thực hiện gần như trọn vẹn những yêu cầu đối với công việc trích xuất văn bản và hình ảnh từ file PDF.
- Bảo mật sản phẩm bằng cách đăng nhập thông qua nhận diện khuôn mặt.

4.1.2. *Nhược điểm*

- Giao diện đẹp chưa được chú trọng.
- Chương trình còn nhiều hạn chế, có xảy ra một số lỗi.
- Dữ liệu còn ít, không được đa dạng.
- Đối với công việc trích xuất văn bản, thuật toán chưa thực sự tối ưu. Nếu muốn trích xuất văn bản mã Unicode thì không thể trích xuất theo định dạng, còn nếu có thể trích xuất văn bản theo định dạng thì không thể trích xuất văn bản mã Unicode hay trích xuất toàn bộ văn bản.
- Đối với công việc trích xuất hình ảnh, kết quả chưa thực sự tốt, nhận diện hình ảnh còn bị sai. Hình ảnh được trích xuất phụ thuộc khá nhiều vào file PDF đầu vào.

4.2. Hướng phát triển sản phẩm

- Chúng em sẽ sửa chữa các lỗi đang mắc phải.
- Xây dựng một quy trình đơn giản hơn.
- Xây dựng giao diện ưa nhìn, thân thiện với người dùng hơn.
- Phát triển sản phẩm và đưa ra sử dụng thực tế.
- Cải tiến thuật toán cho hiệu quả đạt được tốt hơn.

TÀI LIỆU THAM KHẢO

[1] https://quantrimang.com

[2] https://vi.wikipedia.org

[3] https://stackoverflow.com

[4] https://www.tutorialspoint.com

[5] https://en.wikipedia.org

[6] https://tantaingo.wordpress.com

[7] https://www.e-iceblue.com