|  |
| --- |
|  |

|  |
| --- |
| KHOA CÔNG NGHỆ THÔNG TIN – TRƯỜNG ĐH KHOA HỌC TỰ NHIÊN  **QUẢN LÝ QUY TRÌNH PHẦN MỀM** |
| **SNAPTRANS** |
| **THIẾT KẾ HỆ THỐNG THUẬT TOÁN** |

Phiên bản 1.2

Ngày 13/10/2015

**GV Lý thuyết:** Ngô Huy Biên

**Nhóm thực hiện: 7**

lịch sử thay đổi dữ liệu

|  |  |  |  |
| --- | --- | --- | --- |
| **Phiên bản** | **Tác giả** | **Ngày** | **Lý do** |
| 1.0 | 1212473 – Nguyễn Anh Tuấn | 06/10/2015 | Phác thảo tài liệu |
| 1212481 – Phan Thanh Tuấn | Phác thảo tài liệu |
| 1212484 – Trần Anh Tuấn | Phác thảo tài liệu |
| 1212515 – Nguyễn Thanh Vọng | Phác thảo tài liệu |
| 1.1 | 1212430 – Phạm Thị Huyền Trang | 07/10/2015 | Kiểm lỗi chính tả, định dạng văn bản. sửa một vài nội dung |
| 1.2 | 1212481 – Phan Thanh Tuấn | 13/10/2015 | Cập nhật giao diện, thêm thiết kế các class |

# 

**Mục lục**

[I. Giới thiệu 3](#_Toc437182222)

[**1.** **Mục đích của tài liệu** 3](#_Toc437182223)

[**2.** **Đối tượng của tài liệu** 3](#_Toc437182224)

[**3.** **Tác giả** 3](#_Toc437182225)

[II. Thiết kế kiến trúc phần mềm 4](#_Toc437182226)

[III. Thiết kế dữ liệu 5](#_Toc437182227)

[**1.** **Vấn đề** 5](#_Toc437182228)

[**2.** **Giải pháp đưa ra** 5](#_Toc437182229)

[**3.** **Sơ lược về format Dict** 5](#_Toc437182230)

[IV. Thiết kế giao diện 7](#_Toc437182231)

[**1.** **Sơ đồ màn hình** 7](#_Toc437182232)

[**2.** **Đặc tả giao diện** 8](#_Toc437182233)

[V. Thiết kế thuật toán 14](#_Toc437182239)

[**1.** **Bài toán OCR** 14](#_Toc437182240)

[**2.** **Bài toán SMT** 15](#_Toc437182243)

[**3.** **Bài toán đa tiến trình (Multi - Threading)** 16](#_Toc437182246)

[VI. Thiết kế các lớp xử lý 16](#_Toc437182249)

[VII. Thuật ngữ sử dụng 16](#_Toc437182250)

[VIII. Tài liệu tham khảo 16](#_Toc437182251)

# 

# Giới thiệu

## **Mục đích của tài liệu**

Tài liệu này nhằm mục đích mô tả và phân tích một cách đầy đủ, chi tiết các vấn đề thiết kế hệ thống phần mềm **Snaptrans.** Thông qua việc nêu rõ các phần cụ thể từ thiết kế thuật toán, xây dựng kiến trúc, cơ sở dữ liệu đến tạo đối tượng lập trình, thiết kế giao diện. Giúp khách hàng có một cách nhìn chi tiết, cụ thể về phương hướng và chiến lược xây dựng tiến đến hoàn thiện sản phẩm của đội ngũ phát triển.

Tóm lược, tài liệu gồm các mục tiêu sau:

* Mô tả thiết kế kiến trúc
* Mô tả các lớp đối tượng sử dụng
* Cách xây dựng và truy xuất cơ sở dữ liệu
* Trình bày giao diện thiết kế
* Các thuật toán sử dụng

## **Đối tượng của tài liệu**

*Đối tượng người đọc của tài liệu này gồm có:*

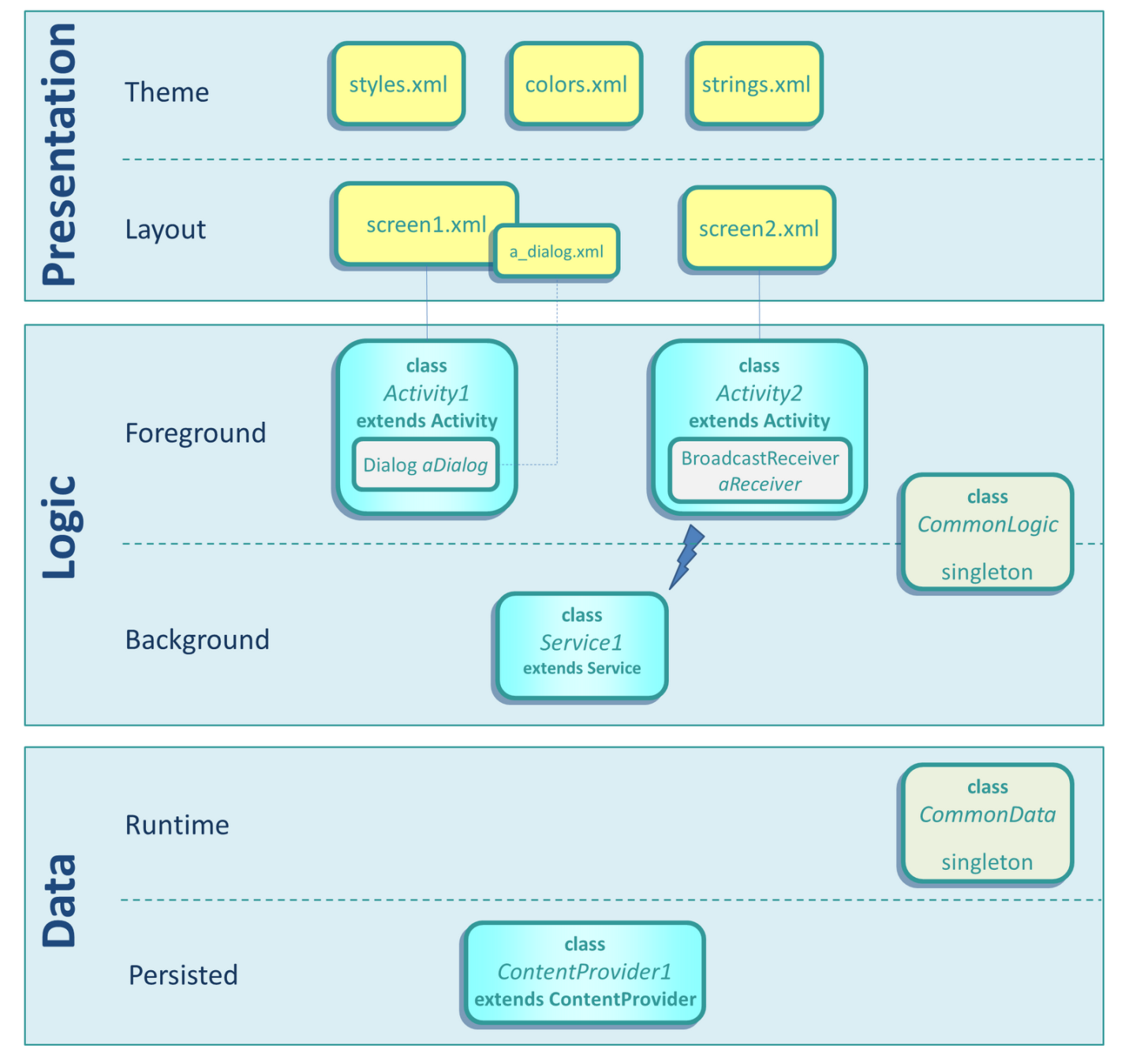
* Khách hàng (thầy Ngô Huy Biên)
* Đội ngũ phát triển sản phẩm Snaptrans, bao gồm 10 thành viên của nhóm 7 môn QLQTPM

## **Tác giả**

*Bộ phận Developer của dự án bao gồm: 1212473 (Nguyễn Anh Tuấn), 1212481 (Phan Thanh Tuấn), 1212484 (Trần Anh Tuấn), 1212515 (Nguyễn Thanh Vọng).*

# Thiết kế kiến trúc phần mềm

Phần mềm được thiết kế theo kiến trúc của một ứng dụng Android



# Thiết kế dữ liệu

## **Vấn đề**

* Yêu cầu xây dựng một hệ thống dữ liệu từ điển.
* Chương trình sử dụng offline và dữ liệu sử dụng gặp phải vấn đề quyền tác giả.
* Việc xây dựng CSDL cho từ điển phải đảm bảo được khả năng truy cập nhanh bởi dữ liệu của từ điển thường khá lớn, lên tới hàng chục nghìn từ.

## **Giải pháp đưa ra**

* Sử dụng dữ liệu từ điển miễn phí được cung cấp bởi tác giả [Hồ Ngọc Đức](http://www.informatik.uni-leipzig.de/~duc/) (loại dữ liệu từ điển sử dụng: [Anh-Việt](http://www.informatik.uni-leipzig.de/~duc/Dict/install.html))
* Sử dụng định dạng từ điển do DICT.ORG (www.dict.org) đã xây dựng, định dạng này đã được dùng để xây dựng những bộ từ điển khá lớn.

## **Sơ lược về format Dict**

Toàn bộ CSDL được chứa trong 2 file, một file chứa nghĩa của từ và một file index. File index bao gồm tên từ, vị trí nghĩa của từ bắt đầu trong file chứa nghĩa và độ dài của nghĩa. Vị trí bắt đầu và độ dài của nghĩa được mã hoá theo cách như sau: Sử dụng 64 chữ cái: *ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwx yz0123456789+/*

chữ cái A tương đương số 0, chữ cái B tương đương số 1 v.v... Giữa từ, vị trí bắt đầu và độ dài nghĩa phân cách nhau bởi ký tự tab (ASCII 9). Mỗi dòng trong file index chứa dữ liệu của một từ. Các dòng phân cách nhau bởi ký tự xuống dòng (ASCII 10).

Ví dụ trong file index của từ điển Đức-Việt có một dòng như sau:

*Abdeckung kbpP D3*

Như vậy nghĩa của từ *Abdeckung* trong file chứa nghĩa sẽ bắt đầu tại offset *kbpP* (theo mã 64 ký tự) và có độ dài là *D3*.

Việc chuyển từ mã cơ số 64 về cơ số 10 được thực hiện như sau:

Đối với vị trí bắt đầu: *kbpP*. Ta có k (ở cơ số 64) = 36 (ở cơ số 10), b = 27, p = 41, P = 15. Như vậy chuyển sang cơ số 10, mã *kbpP* có giá trị là: 36\*643 + 27\*642 + 41\*641 + 15\*640 = 9550415

Đối với độ dài nghĩa: *D3*. Ta có D = 3, 3 = 55. Như vậy chuyển sang cơ số 10, mã *D3* ở cơ số 64 có giá trị là: 247.

File index được sắp xếp để giảm bớt thời gian tìm kiếm. Việc mã hoá theo cơ số 64 như trên giúp cho kích thước file index giảm xuống rất nhiều so với không mã hóa.

Còn cấu trúc của file chứa nghĩa gồm các phần như sau:

*@headword*

*\* tu loai (noun, verb...)*

*- dinh nghia 1*

*= cau vi du cho dinh nghia 1 + nghia cua cau do*

*- dinh nghia 2*

*= cau vi du cho dinh nghia 2 + nghia cua cau do*

*\* tu loai*

*- dinh nghia 3*

Nghĩa của mỗi từ gồm một phần như trên, các nghĩa của mỗi từ nối tiếp nhau liên tục.

***Hàm chuyển từ mã cơ số 64 sang cơ số 10 bằng ngôn ngữ java:***

*public Int GetDemicalValue (CString str){*

*CString base64 = 'ABCDEFGHIJKLM-NOPQRSTUVWXYZabcdefghijklmnopqrstu-vwxyz0123456789+/';*

*int decValue = 0;*

*int len = str.GetLength();*

*for (int i = 0; i<len; i++){*

*int pos = base64.Find(str.GetAt(i),0);*

*decValue += (int)pow(64,len-i- 1)\*pos;*

*}*

*return decValue;*

*}*

# Thiết kế giao diện

## **Sơ đồ màn hình**

OCR

Camera

Translate

Settings

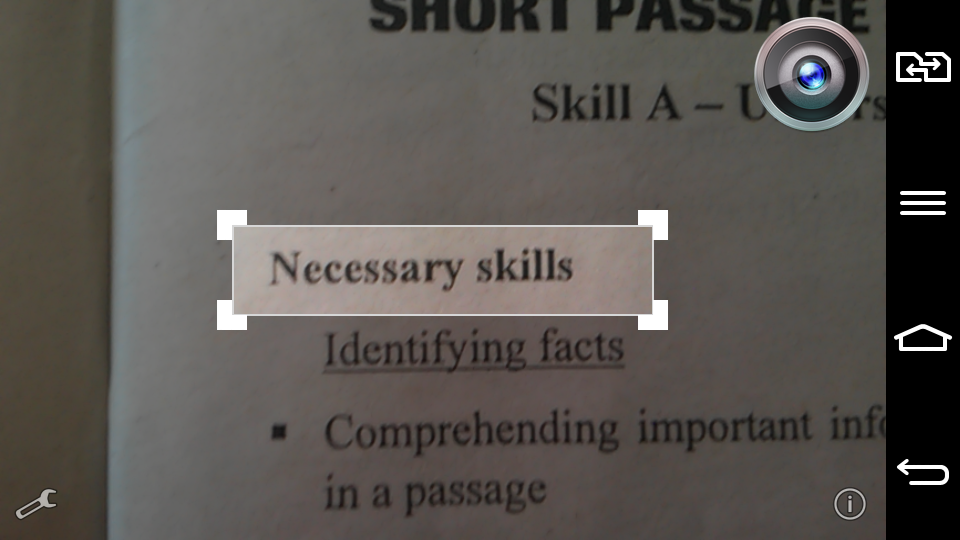
About

Share

## **Đặc tả giao diện**

1

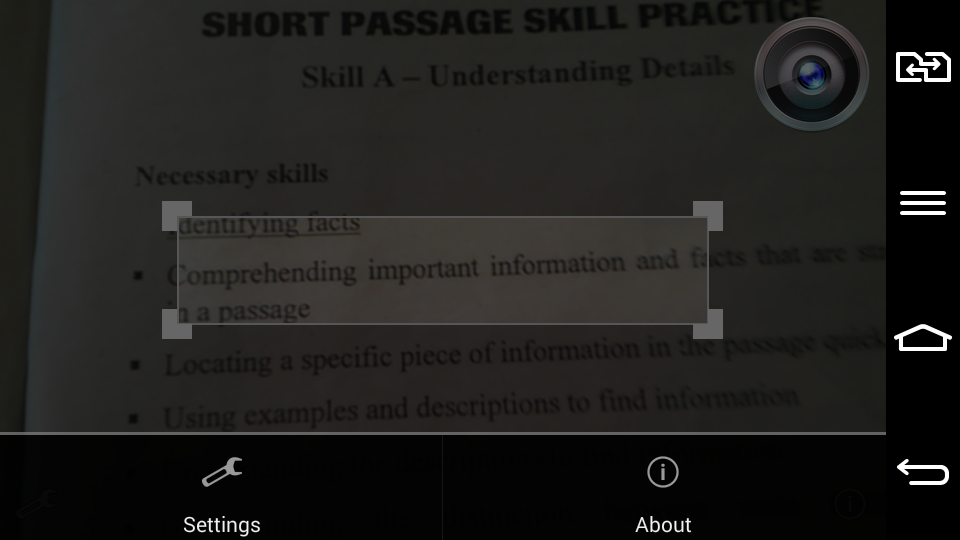
### **Camera**



2

4

3

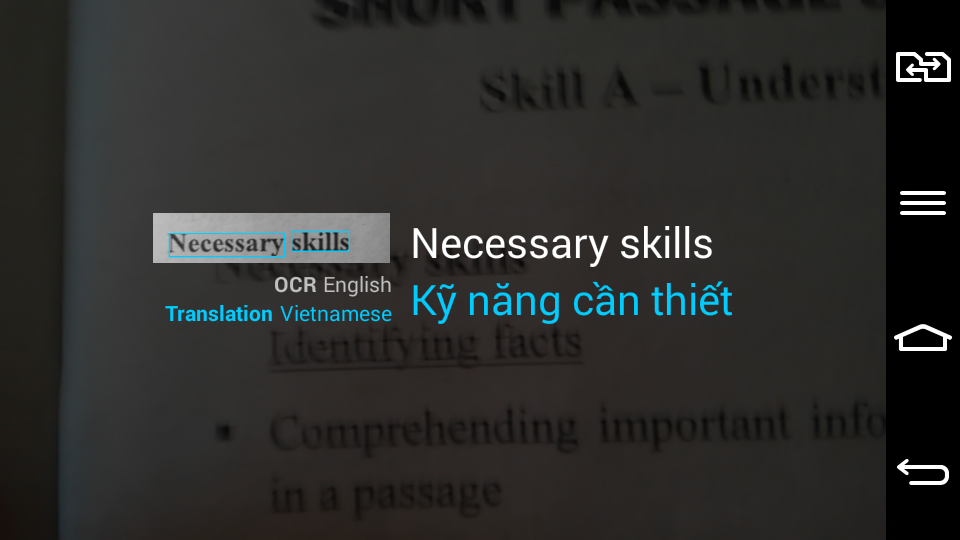


3

4

|  |  |  |
| --- | --- | --- |
|  | **Ràng buộc** | **Mô tả** |
| 1 | -Loại: Button | Nhấn vào để hiển thị kí tự nhận dạng được và dịch sang ngôn ngữ đích |
| 2 | -Loại: Rect | Có thể tùy chỉnh hình chữ nhật để được vùng nhận dạng ảnh theo ý muốn |
| 3 | -Loại: Button | Nhấn vào chuyển qua màn hình Settings |
| 4 | -Loại: Button | Nhấn vào chuyển qua màn hình About |

### **Translate và OCR**



3

2

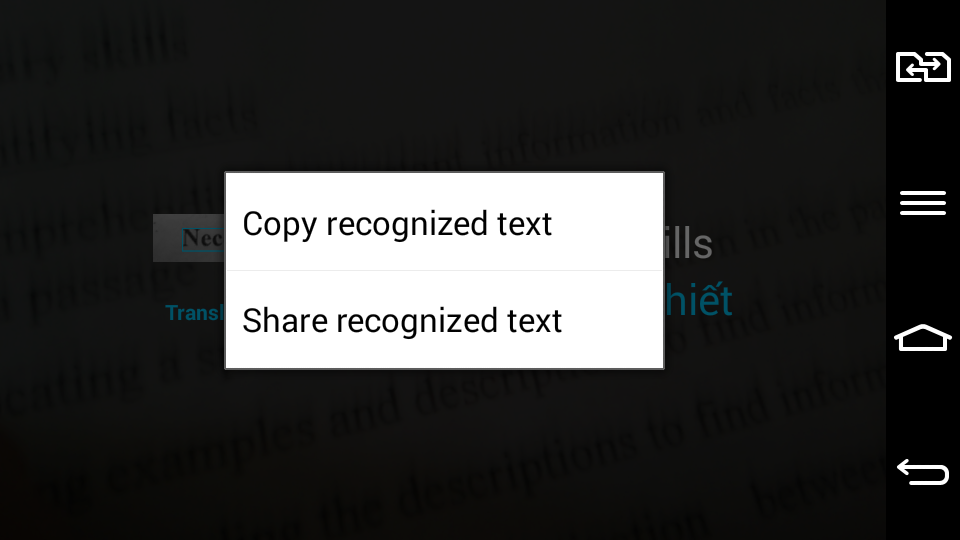
5

4

1

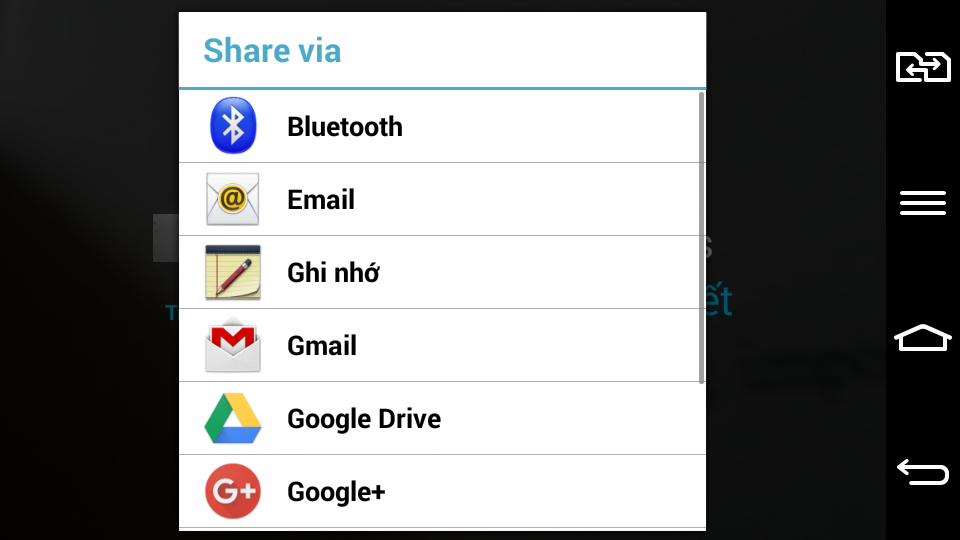
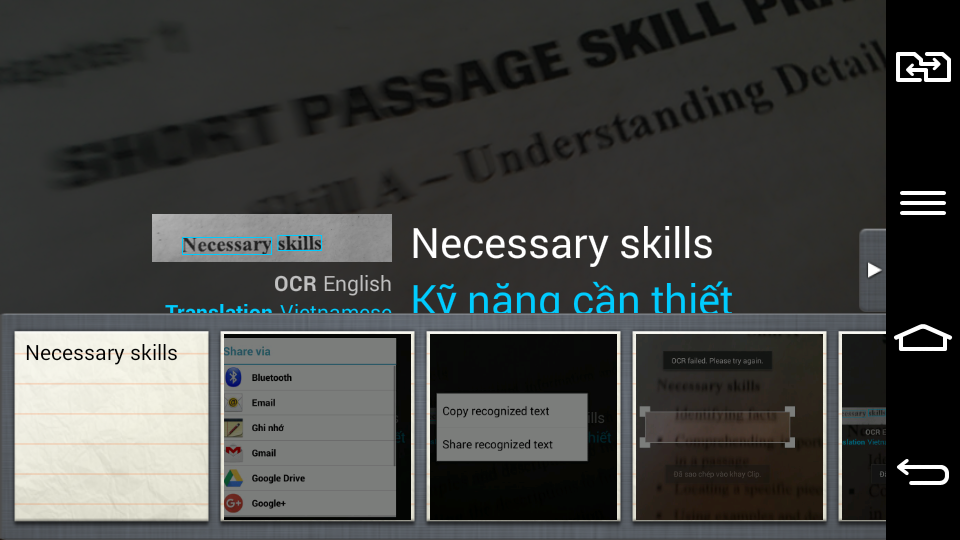
|  |  |  |
| --- | --- | --- |
|  | **Ràng buộc** | **Mô tả** |
| 1 | -Loại: Image | Ảnh chụp đoạn văn bản cần nhận diện |
| 2 | -Loại: TextView | Tên ngôn ngữ nguồn cần nhận dạng |
| 3 | -Loại: TextView | Tên ngôn ngữ đích cần dịch |
| 4 | -Loại: TextView | Kết quả nhận dạng từ ảnh nguồn |
| 5 | -Loại: TextView | Kết quả dịch ra ngôn ngữ đích từ kí tự nhận dạng được |

### **Share và Coppy**



2

1

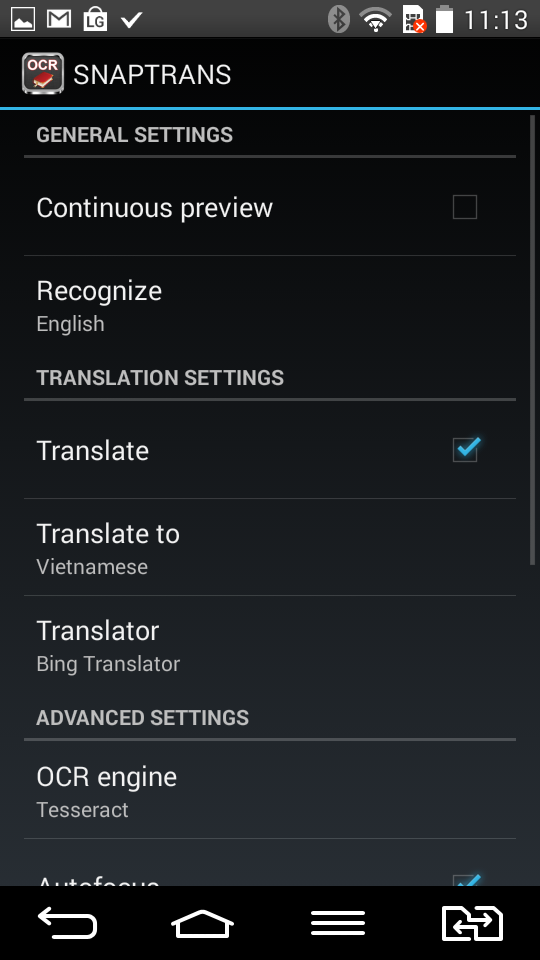


2

1

|  |  |  |
| --- | --- | --- |
|  | **Ràng buộc** | **Mô tả** |
| 1 | -Loại: ContextMenu | Nhấn vào để sao chép văn bản nhận dạng hoặc văn bản dịch vào clipboard |
| 2 | -Loại: ContextMenu | Nhấn vào để chia sẻ văn bản nhận dạng hoặc văn bản dịch thông qua các ứng dụng khác. |

### **Settings**



6

5

4

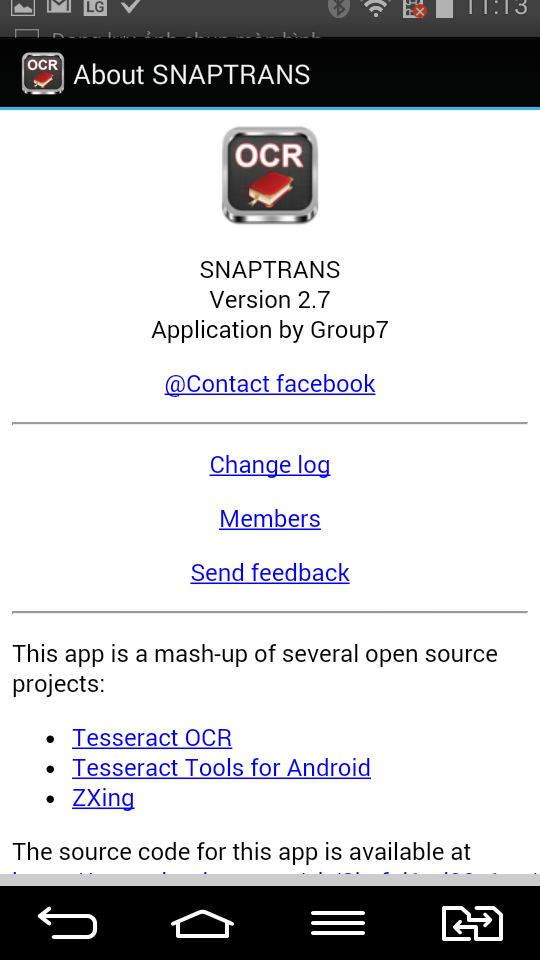
3

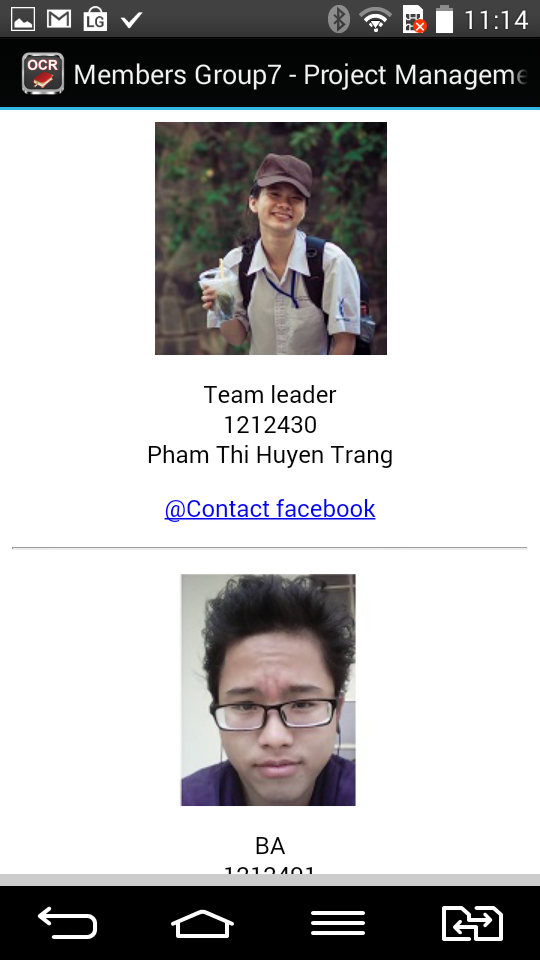
1

2

|  |  |  |
| --- | --- | --- |
|  | **Ràng buộc** | **Mô tả** |
| 1 | -Loại: CheckBox | Bật hoặc tắt TextView cập nhật hiển thị kí tự nhận dạng và TextView theo dõi quá trình nhận dạng trong giao diện camera |
| 2 | -Loại: List | Nhấn vào thay đổi ngôn ngữ nguồn cần nhận dạng |
| 3 | -Loại: CheckBox | Bật hoặc tắt chức năng dịch ra ngôn ngữ đích |
| 4 | -Loại: List | Nhấn vào thay đổi ngôn ngữ đích cần dịch |
| 5 | -Loại: List | Nhấn vào thay đổi công cụ dịch |
| 6 | -Loại: List | Nhấn vào thay đổi tập huấn luyện nhận dạng |

### **About**





# Thiết kế thuật toán

## **Bài toán OCR**

### **Vấn đề**

* Nhận diện chữ viết từ camera điện thoại (yêu cầu độ chính xác về nhận diện các từ là 80%)
* Chương trình sử dụng offline

### **Giải pháp đưa ra**

* Sử dụng thư viện OCR mã nguồn mở có hỗ trợ cho android: [Tesseract](https://code.google.com/p/tesseract-ocr/)
* [Tập dữ liệu huấn luyện](https://code.google.com/p/tesseract-ocr/downloads/list) cho một ngôn ngữ (dữ liệu tiếng Anh được sử dụng)
* Cài đặt [android NDK](https://developer.android.com/tools/sdk/ndk/index.html) cho Android Studio
* Ví dụ mã nguồn Java hàm phát hiện kí tự số và trả về dạng chuỗi:

*public String detectNumber(Bitmap bitmap) {*

*TessDataManager.initTessTrainedData(context);*

*TessBaseAPI tessBaseAPI = new TessBaseAPI();*

*String path = "/mnt/sdcard/packagename/tessdata/eng.traineddata";*

*tessBaseAPI.setDebug(true);*

*tessBaseAPI.init(path, "eng"); //Init the Tess with the trained data file, with english language*

*//For example if we want to only detect numbers*

*tessBaseAPI.setVariable(TessBaseAPI.VAR\_CHAR\_WHITELIST, "1234567890");*

*tessBaseAPI.setVariable(TessBaseAPI.VAR\_CHAR\_BLACKLIST, "!@#$%^&\*()\_+=-qwertyuiop[]}{POIU" +*

*"YTREWQasdASDfghFGHjklJKLl;L:'\"\\|~`xcvXCVbnmBNM,./<>?");*

*tessBaseAPI.setImage(bitmap);*

*String text = tessBaseAPI.getUTF8Text();*

*Log.d(TAG, "Got data: " + result);*

*tessBaseAPI.end();*

*return text;*

*}*

## **Bài toán SMT**

### **Vấn đề**

* Dịch tất cả các chữ trong văn bản từ một ngôn ngữ nguồn sang ngôn ngữ đích.
* Mỗi câu văn dịch được trong ngôn ngữ đích giống với ngôn ngữ tự nhiên và sát với ý câu văn nguồn.

### **Giải pháp đưa ra**

* Chỉ giải quyết được 50% yêu cầu (dịch từ với từ – word by word)
* Vì thời gian phát triển sản phẩm quá ngắn, không đủ để nhóm nghiên cứu sâu vào dịch máy thông kê. Nếu thời gian cho phép, trong quá trình hoàn thiện sản phẩm có thể áp dụng mô hình dịch dựa vào cụm từ sử dụng phần mềm mã nguồn mở [Moses](http://www.statmt.org/moses/).

## **Bài toán đa tiến trình (Multi - Threading)**

### **Vấn đề**

Ứng dụng không bị treo trong mọi trường hợp như: khi khởi động, khi chụp ảnh, khi dịch, khi tra từ….

### **Giải pháp đưa ra**

dùng *sendMessage* hoặc *Post* của *Handler* class để xử lý đa tiến trình.

# Thiết kế các lớp xử lý

[Xem chi tiết tại website của nhóm](http://documentgroup7.byethost13.com/)

<http://documentgroup7.byethost13.com/>

# Thuật ngữ sử dụng

|  |  |
| --- | --- |
| **Thuật ngữ** | **Ý nghĩa** |
| QLQTPM | Quản lý quy trình phần mềm |
| SDS | Software Design Specification |
| OCR | Optical Character Recognition |
| CSDL | Cơ sở dữ liệu |
| SMT | Statistical Machine Translation |

# Tài liệu tham khảo

<http://www.toptal.com/freelance/why-design-documents-matter>

<http://www.cs.iit.edu/~oaldawud/CS487/project/software_design_specification.htm>

<https://code.google.com/p/tesseract-ocr/>

<http://www.informatik.uni-leipzig.de/~duc/>

<http://www.informatik.uni-leipzig.de/~duc/Dict/install.html>